2018

STATISTICAL METHODS FOR DETECTING CAUSAL RARE VARIANTS AND ANALYZING MULTIPLE PHENOTYPES

Xinlan Yang
Michigan Technological University, xinlany@mtu.edu

STATISTICAL METHODS FOR DETECTING CAUSAL RARE VARIANTS AND ANALYZING MULTIPLE PHENOTYPES

By

Xinlan Yang

A DISSERTATION

Submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

In Mathematical Sciences

MICHIGAN TECHNOLOGICAL UNIVERSITY

2018

This dissertation has been approved in partial fulfillment of the requirements for the Degree of DOCTOR OF PHILOSOPHY in Mathematical Sciences.

Department of Mathematical Sciences

Dissertation Advisor:	*Qiuying Sha*
Committee Member:	*Shuanglin Zhang*
Committee Member:	*Kui Zhang*
Committee Member:	*Jingfeng Jiang*
Department Chair:	*Mark S. Gockenbach*

Contents

List of tables

List of figures

Preface

This dissertation is submitted for the degree of Doctor of Philosophy at Michigan Technological University. The research described herein was conducted under the supervision of Prof. Qiuying Sha in the Department of Mathematical Sciences, Michigan Technological University, between September 2013 and March 2018.

This work is to the best of my knowledge original, except where references are made to previous work. Part of this work contains previously published material. The title of Chapter 1 is *Detecting Association of Rare and Common Variants based on Cross-Validation Prediction Error* and it was published in the Genetic Epidemiology (Yang X, Wang S, Zhang S, Sha Q, 2017 Apr; 41(3):233-243). The overall study was designed by Qiuying Sha. Xinlan Yang, Shuaichen Wang and Shuanglin Zhang conducted the statistical analyses. Xinaln Yang, Shuanglin Zhang, and Qiuying Sha drafted the manuscript. Chapter 2 is in preparation for future publication and contains its introduction, methods, simulation study, results, real data analysis and discussion sections. Shuanglin Zhang, and Qiuying Sha designed research, Xinlan Yang and Shuanglin Zhang performed statistical analysis, and Xinlan Yang, Shuanglin Zhang, and Qiuying Sha wrote the manuscript.

Acknowledgements

During the 5 years of study at Michigan Technological University, I have acquired knowledge and experiences of data analysis not only from textbooks and practices, but also from many knowledgeable and respectable professionals. Therefore, I would like to express my deep appreciations to those significant ones who helped me the most.

Firstly, I would like to express my gratitude to the committee for their support. It is a great honor to have these outstanding scholars be part of the committee. I would like to give my thanks to Dr. Shuanglin Zhang for his inspiring tips and advice on my research topics; appreciations to Dr. Kui Zhang and Dr. Jingfeng Jiang who carefully examined my dissertation and gave their wise suggestions; the most special thanks to Dr. Qiuying Sha, my advisor, who offered the most support in my whole PhD program. More than a successful professor in career, she's also a charismatic person, an academic role model, and a caring parent figure to all her students. I will always be inspired by her profound insights and unremitting efforts in statistical genetics.

Secondly, I'd like to thank professors who gave instructions, along with my fellow students and researchers: Shuaichen Wang, Dr. Xiao Zhang and Dr. Mark S. Gockenbach.

Finally, I dedicate this dissertation to my family. Special thanks are due to my parents Mrs. Caiyun Chen and Mr. Ji Yang for their boundless love and dedication to me. Without their unconditional support and continuous expression of pride, I would never have made my achievements during the graduate study.

Abstract

This dissertation includes two papers with each distributed in one chapter.

To date, genome-wide association studies (GWAS) have identified a large number of common variants that are associated with complex diseases successfully. However, the common variants identified by GWAS only account for a small proportion of trait heritability. Many studies showed that rare variants could explain parts of the missing heritability. Since the well-developed common variant detecting methods are underpowered for rare variant association tests unless sample sizes or effect sizes are very large, investigation the roles of rare variants in complex diseases presents substantial challenges. In chapter 1, we proposed novel statistical tests to test the association between rare and common variants in a genomic region and a complex trait of interest based on cross-validation prediction error. we first proposed a prediction error method (PE) based on Ridge regression. Based on PE, we also proposed another two tests PE-WS and PE-TOW by testing a weighted combination of variants with two different weighting schemes. Using extensive simulation studies, we showed that PE-TOW and PE-WS are consistently more powerful than TOW and WS, respectively, and PE is the most powerful test when causal variants contain both common and rare variants.

In genome-wide association studies (GWAS), the joint analysis of multiple phenotypes could have increased power over analyzing each phenotype individually. With this motivation, several methods that jointly analyze multiple phenotypes have been developed, such as O'Brien's method, Trait-based Association Test that uses Extended

Simes procedure (TATES), MAONVA and MultiPhen. However, the performance of these methods under a wide range of scenarios is not consistent: one test may be powerful in some situations, but not in the others. Thus, one challenge in joint analysis of multiple phenotypes is to construct a test that could maintain good performance across different scenarios. In chapter 2, we developed a novel statistical method to test the association between a genetic variant and multiple phenotypes based on cross-validation prediction error. Extensive simulations were conducted to evaluate the type I error rates and to compare the power performance of the PE method with various existing methods. Simulation studies showed that the PE method controls the type I error rates very well and has consistently higher power than the tests we compared in all the scenarios.

1 Chapter 1

Detecting Association of Rare and Common Variants based on Cross-Validation Prediction Error

Despite the extensive discovery of disease-associated common variants, much of the genetic contribution to complex traits remains unexplained. Rare variants may explain additional disease risk or trait variability. Although sequencing technology provides a supreme opportunity to investigate the roles of rare variants in complex diseases, detection of these variants in sequencing-based association studies presents substantial challenges. In this article, we propose novel statistical tests to test the association between rare and common variants in a genomic region and a complex trait of interest based on cross-validation prediction error. We first propose a prediction error method (PE) based on Ridge regression. Based on PE, we also propose another two tests PE-WS and PE-TOW by testing a weighted combination of variants with two different weighting schemes. PE-WS is the prediction error version of the test based on the weighted sum statistic (WS) and PE-TOW is the prediction error version of the test based on the optimally weighted combination of variants (TOW). Using extensive simulation studies, we are able to show that (1) PE-TOW and PE-WS are consistently more powerful than TOW and WS, respectively, and (2) PE is the most powerful test when causal variants contain both common and rare variants.

1.1 Introduction

The main purpose of genome-wide association studies (GWAS) is to detect common variants by indirect mapping methods. GWAS have identified a large number of common variants that are associated with complex diseases successfully [Bodmer and Bonilla, 2008; Lango Allen *et al*., 2010; Ng *et al*., 2009; Pritchard, 2001; Pritchard and Cox, 2002; Stratton and Rahman, 2008; Teer and Mullikin, 2010; Walsh and King, 2007]. However, the common variants identified by GWAS only account for a small fraction of trait heritability [McCarthy *et al*., 2008], parts of the missing heritability could be caused by rare variants [Cohen *et al*., 2006; Ji *et al*., 2008; Manolio *et al*., 2009; Marini *et al*., 2008; Zhu *et al*., 2010]. In rare variant association studies, instead of indirect association mapping method, all rare variants need to be tested directly. The new sequencing technology allows sequencing of exome-wide and whole-genome of a large amount of individuals [Hodges *et al*., 2007], which makes directly test for rare variants possible [Andre's *et al*., 2007]. Current exome-wide and whole-genome sequencing studies have successfully detected many rare variants responsible for many complex traits, such as LDL Cholesterol [Lange *et al*., 2014], bone mineral density [Huang *et al*., 2015], thyroid function [Zheng *et al*., 2015], circulating lipid levels [Taylor *et al*., 2015], and other traits [Walter *et al*., 2015].

There is an increasing number of researchers who are interested in rare variants association studies [Cohen *et al*., 2004; Ji *et al*., 2008; Ahituv *et al*., 2007; Romeo *et al*., 2007; Romeo *et al*., 2009]. Since the well-developed common variant detecting methods

are underpowered for rare variant association tests unless sample sizes or effect sizes are very large, several new methods for rare variant association studies are proposed recently. These methods include burden tests, quadratic tests, and robust tests. Burden tests include the cohort allelic sums test (CAST) [Morgenthaler and Thilly, 2007], the combined multivariate and collapsing (CMC) method [Li and Leal, 2008], the weighted sum statistic (WS) [Madsen and Browning, 2009], and variable threshold (VT) method [Price *et al*., 2010]. Burden tests collapse rare variants in a genomic region into a single burden variable and then regress the phenotype on the burden variable to test for the cumulative effects of rare variants in the region [Lee *et al*., 2012]. These tests implicitly assume that all rare variants are causal and that the directions of the effects are all the same. Quadratic tests include tests with statistics of quadratic forms of the score vector such as the sequence kernel association test (SKAT) [Wu *et al*., 2011], the sequence kernel association test for the combined effect of rare and common variants (SKAT-C) [Ionita-Laza *et al*., 2013], the test for optimally weighted combination of variants (TOW) [Sha *et al*., 2012], as well as adaptive weighting methods such as data-adaptive sum (aSUM) [Han and Pan, 2010], adaptive weighting (AW) methods [Sha *et al*., 2013], and methods proposed by Hoffmann *et al*. [2010], Lin and Tang [2011], and Yi and Zhi [2011]. Quadratic tests are robust to the directions of the effects of causal variants and are less affected by neutral variants than burden tests. Burden tests can only outperform quadratic tests when most of rare variants are causal and the directions of the effects of causal variants are all the same. Robust tests include methods proposed by Derkach *et al*. [2012], Lee *et al*. [2012], Greco *et al*. [2016], and Sha and Zhang [2014]. Robust tests combine

information from burden tests, quadratic tests, and possibly other tests aiming to have advantages of burden, quadratic, and possibly other tests.

In this paper, we develop novel statistical methods to test the association between common and rare variants in a genomic region and a complex trait of interest based on cross-validation prediction error. We first propose a prediction error method (PE) based on Ridge regression. Based on PE, we also propose another two tests PE-WS and PE-TOW by testing a weighted combination of variants with two different weighting schemes, the weights suggested by Madsen and Browning [2009] and the optimal weighting scheme developed by Sha *et al*. [2012]. By extensive simulation studies, we show that (1) the prediction error versions of TOW and WS (PE-TOW and PE-WS) are consistently more powerful than TOW and WS, respectively, and (2) PE is the most powerful test when the causal variants contain both common and rare variants.

1.2 Method

Prediction Error Model

Consider a sample of n unrelated individuals. Each individual has been genotyped at M variants in a genomic region. Denote y_i as the value of a quantitative trait of the i^{th} individual and denote $g_i = (g_{i1},...,g_{iM})^T$ as the genotypic scores of the i^{th} individual at M variants, where $g_{im} \in \{0, 1, 2\}$ is the number of minor alleles the i^{th} individual has at the m^{th} variant. We assume that there are no covariates. If there are p covariates, $z_{i1},\ldots,z_{ip}$, for the i^{th} individual, we adjust genotypes and trait values for the covariates using the method applied by Price *et al*. [2006] and Sha *et al*. [2012], that is, adjusting both genotypes and trait values for the covariates through linear models

$y_i = \alpha_0 + \alpha_1 z_{i1} + ... + \alpha_p z_{ip} + \varepsilon_i$ and $g_{im} = \alpha_{0m} + \alpha_{1m} z_{i1} + ... + \alpha_{pm} z_{ip} + \tau_{im}$.

Our working model is

$$y_i = \beta_0 + \beta_1 g_{i1} + \cdots + \beta_M g_{iM} + \varepsilon_i . \quad (1)$$

To test association under our working model (1), we test the null hypothesis $H_0 : \beta_1 = \cdots = \beta_M = 0$.

In the k-fold cross-validation, we divide the data into k equal parts, then use each of the k parts as the testing set and the other k-1 parts as the training set. We use the training set

to estimate $\beta = (\beta_0, \ldots, \beta_M)^T$ in equation (1), and use the prediction equation $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 g_{i1} + \cdots + \hat{\beta}_M g_{iM}$ to predict the trait values in the testing set. Since the genotype data of rare variants are sparse, the smaller the training set is, the more likely the problem of singular of the design matrix will be. Thus, we should try to use the training set as large as possible. In the k-fold cross-validation, the leave-one-out cross-validation (LOOCV) (k equals n) gives the largest training sets. Furthermore, the LOOCV prediction error has a closed-form formula [James *et al*., 2013] (also see Appendix A for $\lambda = 0$). Therefore, our proposed tests are based on LOOCV.

In this paper, we construct a novel statistical test to test the association between genotypes of common and rare variants in a genomic region and a complex trait of interest based on the LOOCV prediction error. We propose to use the LOOCV prediction error under model (1) as a test statistic. Let $\hat{y}_{ci}$ denote the LOOCV predicted value of y_i under model (1). Then, the statistic can be written as

$$T = \sum_{i=1}^{n} (y_i - \hat{y}_{ci})^2 . \qquad (2)$$

Note that T is the LOOCV prediction error. Thus, low values of T would imply significance.

Ridge regression

For rare variants, one drawback of the aforementioned LOOCV procedure is that some columns of the design matrix may have all zeros, if we leave one individual out. When the design matrix is not full rank or columns of the design matrix are highly correlated, we can use penalized regressions, such as Ridge regression [Halawa and Bassiouni, 2000; Hoerl *et al*., 1975] and Lasso regression [Yuan and Lin, 2006; Meier *et al*., 2008; Tibshirani, 1996] among others. Penalized regressions have been applied to the analysis of genetic data [Cule *et al*., 2011; Warren *et al*., 2014; Ayers and Cordell, 2013; Ayers and Cordell, 2010; Cule and De Iorio, 2013; Malo *et al*., 2008]. In this paper, we propose to use Ridge regression. Ridge regression penalizes the size of the regression coefficients. Let $x_i = (1, g_{i1}, ..., g_{iM})^T$ and $\beta = (\beta_0, \ldots, \beta_M)^T$. In the regression model $y_i = x_i^T \beta + \varepsilon_i,\ i = 1, 2, ..., n$, the Ridge regression estimator $\hat{\beta}$ is defined as the value of β that minimizes $\sum_i (y_i - x_i^T \beta)^2 + \lambda \sum_j \beta_j^2$, where $\lambda \geq 0$ is a tuning parameter. The solution to the Ridge regression is given by $\hat{\beta} = (X^T X + \lambda I)^{-1} X^T y$, where $X = (x_1, \ldots, x_n)^T$. The LOOCV prediction error for Ridge regression also has a closed-form formula (see Appendix A). For Ridge regression, we denote the test statistic given by equation (2) as T_λ. Let p_λ denote the p-value of T_λ, where p_λ is evaluated using equation (4) in the next section and $p_\lambda = p_\lambda^{(0)}$. We define the LOOCV Prediction Error test statistic (PE) as

$$T_{PE} = \min_\lambda p_\lambda . \tag{3}$$

In this study, we use a simple method to evaluate the minimization. We divide the interval $[0,\infty)$ into subintervals $0 \le \lambda_1 < ... < \lambda_{K-1} < \lambda_K < \infty$. In the simulation studies (see later), we used $K = 10$ and $(\log \lambda_1, \ldots, \log \lambda_{10}) = (1, \ldots, 10)$. Then, $T_{PE} = \min_\lambda p_\lambda = \min_{1 \le k \le K} p_{\lambda_k}$

.

We use a permutation procedure to evaluate the p-value of T_{PE}. Intuitively, two layers of permutations are needed to estimate p_{λ_k} and the overall p-value for the test statistic T_{PE} . Ge *et al*. [2003] proposed that one layer of permutation can be used to estimate p_{λ_k} and the overall p-value for the test statistic T_{PE}. Here, we use the permutation procedure of Ge *et al*. to estimate p_{λ_k} and the overall p-value for the test statistic T_{PE}. In each permutation, we randomly shuffle the trait values. Suppose that we perform B replicates of permutations. Let $T_{\lambda_k}^{(b)}$ denote the values of T_{λ_k} based on the b^{th} permuted data for $b = 0,1,...,B$ and $k = 1, \ldots, K$, where $b = 0$ represents the original data. Then, we transfer $T_{\lambda_k}^{(b)}$ to $p_{\lambda_k}^{(b)}$ by

$$p_{\lambda_k}^{(b)} = \frac{\#\{d : T_{\lambda_k}^{(d)} < T_{\lambda_k}^{(b)} \text{ for } d = 1, \ldots, B\}}{f(b)}, \tag{4}$$

where $f(0) = B$ and $f(b) = B - 1$ for $b = 1, \ldots, B$. Let $p^{(b)} = \min_{1 \le k \le K} p_{\lambda_k}^{(b)}$. Then, the p-value of T_{PE} is given by

$$\frac{\#\{b : p^{(b)} < p^{(0)} \text{ for } b = 1,2,\ldots,B\}}{B}.$$

See Appendix B for a fast algorithm for the permutation procedure.

For testing the effects of common and rare variants, we also propose the following two methods based on the framework of PE. These two methods are to test the effect of a weighted combination of variants with two different weighting schemes:

1. Weighted sum weighting scheme: in this weighting scheme, we replace $g_i = (g_{i1},...,g_{iM})^T$ with $G_i = \sum_{m=1}^{M} w_m g_{im}$, where $w_m = \frac{1}{\sqrt{p_m(1-p_m)}}$ is the weight suggested by Madsen and Browning [2009] and p_m is the minor allele frequency of the m^{th} variant. The test statistic given by equation (3) based on this weighting scheme is called weighted sum method based on prediction error (PE-WS).

2. Optimal weighting scheme: in this weighting scheme, we replace $g_i = (g_{i1},...,g_{iM})^T$ with $G_i = \sum_{m=1}^{M} w_m g_{im}$, where $w_m = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(g_{im} - \bar{g}_m)}{\sum_{i=1}^{n}(g_{im} - \bar{g}_m)^2}$ is the weight suggested by Sha *et al*. [2012]. The test statistic given by equation (3) based on this weighting scheme is called testing an optimally weighted combination of variants based on prediction error (PE-TOW).

We use the same permutation procedure as PE to evaluate the p-values of PE-WS and PE-TOW. See Appendix B for fast algorithms for the permutation procedures of PE-WS and PE-TOW.

Comparison of Tests

We compare the performance of the three proposed tests, PE, PE-WS, and PE-TOW, with that of the weighted sum statistic (WS) [Madsen and Browning, 2009], the sequence kernel association test (SKAT) [Wu *et al*., 2011], the sequence kernel association test for the combined effect of rare and common variants (SKAT-C) [Ionita-Laza *et al*., 2013], and the test for the optimally weighted combination of variants (TOW) [Sha *et al*., 2012].

1.3 Simulation Study

In simulation studies, we generate genotype data using the Genetic Analysis Workshop 17 (GAW17) data. This dataset contains genotypes of 697 unrelated individuals on 3205 genes. Similar to Sha *et al*. [2012], we choose four genes: ELAVL4, MSH4, PDE4B, and ADAMTS4 with 10, 20, 30, and 40 variants, respectively, and then merge the four genes to form a super gene (Sgene) with 100 variants. We generate genotypes based on the genotypes of 697 individuals in the Sgene.

To evaluate type I error, we generate trait values independent of genotypes by using the model:

$$y = 0.5Z_1 + 0.5Z_2 + \varepsilon \tag{5}$$

where Z_1 is a continuous covariate generated from a standard normal distribution, Z_2 is a binary covariate taking values 0 and 1 with a probability of 0.5, and ε follows a standard normal distribution.

To evaluate power, we randomly choose n_c rare variants and one common variant as causal variants and assume that all the n_c rare causal variants have the same heritability. n_r and n_p are the number of risk rare variants and protective rare variants, respectively, then $n_r + n_p = n_c$. Denote x_i^r, x_j^p, and x_c as the genotypes of the i^{th} risk rare variant, the

j^{th} protective rare variant, and the common causal variant, respectively. Then, we generate a quantitative trait by the following model:

$$y = 0.5Z_1 + 0.5Z_2 + \sum_{i=1}^{n_r} \beta_i^r x_i^r - \sum_{j=1}^{n_p} \beta_j^p x_j^p + \beta_c x_c + \varepsilon \quad (6)$$

where Z_1, Z_2 and ε are the same as those in equation (5). In equation (6), β_i^r, β_j^p and β_c are constant coefficients. The values of β_i^r, β_j^p and β_c depend on the total heritability h_{total} and the ratio of the heritability of rare causal variants to the heritability of the common causal variant R. For given h_{total} and R, based on equation (6), we can calculate the heritability of the rare casual variants and the heritability of the common causal variant. From the heritability of the common causal variant, we can calculate β_c. From the heritability of the rare casual variants and the assumption that all the rare causal variants have the same heritability, we can calculate the heritability of each rare causal variant. Then, we can calculate β_i^r and β_j^p. The formulae to calculate the values of β_i^r, β_j^p and β_c are given by $\beta_i^r = \sqrt{\frac{h_{total} R}{\mathrm{var}\left(x_i^r\right) n_c \left(1 - h_{total}\right)\left(1 + R\right)}}$, $\beta_j^p = -\sqrt{\frac{h_{total} R}{\mathrm{var}\left(x_j^p\right) n_c \left(1 - h_{total}\right)\left(1 + R\right)}}$, and $\beta_c = \sqrt{\frac{h_{total}}{\mathrm{var}\left(x_c\right)\left(1 - h_{total}\right)\left(1 + R\right)}}$, respectively.

For power comparisons, we consider two different cases: (1) “Rare” in which all causal variants are rare (minor allele frequency < 0.01) and (2) “Both” in which both rare and common variants contribute to the trait. In each case, we consider two subcases: with

covariates and without covariates. In the subcase of without covariates, Z_1, Z_2 are not included in equation (6).

1.4 Simulation Results

For evaluating the Type I error of the proposed methods (PE, PE-TOW and PW-WS), we consider different disease models (with or without covariates), different significance levels, and different sample sizes. The p-values are calculated using 10,000 permutations. Type I error rates are evaluated using 10,000 replicated samples. For 10,000 replicated samples, the 95% confidence intervals (CIs) for the estimated type I error rates of nominal levels 0.05, 0.01, and 0.001 are (0.046, 0.054), (0.008, 0.012), and (0.00038, 0.00162), respectively. The estimated type I error rates of the three proposed tests are summarized in Table 1. From this table, we can see that most of the estimated type I error rates are within 95% CIs and those type I error rates not within the 95% CIs are very close to the bound of the corresponding 95% CI, which indicates that the proposed methods are valid.

In power comparisons, the p-values of PE, PE-TOW, PE-WS, and TOW are calculated using 1,000 permutations, while the p-values of WS, SKAT, and SKAT-C are calculated by asymptotic distributions. The powers of all of the seven tests are evaluated using 1,000 replicated samples at a significance level of 0.05 (Figures 1-3). For Figure 4, the powers of all of the seven tests are evaluated using 1,000 replicated samples at a significance level of 10^{-6} and p-values of PE-WS, PE-TOW, PE and TOW are evaluated by 10^7 permutations.

Power comparisons of the seven tests (PE, PE-TOW, TOW, PE-WS, WS, SKAT, and SKAT-C) for the power as a function of heritability are given in Figure 1. As shown in Figure 1, (1) PE-WS and PE-TOW are consistently more powerful than WS and TOW, respectively; (2) PE is the most powerful test when the causal variants contain both common and rare variants; and PE is the least powerful test when the causal variants are all rare variants; (3) TOW is more powerful than SKAT when the causal variants are all rare variants (MAF<0.01) and TOW is less powerful than SKAT when the causal variants contain both common and rare variants. The reasons are that (a) TOW and SKAT have different weights, otherwise TOW and SKAT are same and (b) the weights of SKAT are larger than that of TOW only for those variants with MAF in the range (0.01,0.035), and the weights of TOW and SKAT are similar for those variants with MAF>0.035; and (4) SKAT-C is less powerful than SKAT when the causal variants are all rare variants (MAF<0.01) and SKAT-C is more powerful than SKAT when the causal variants contain both common and rare variants.

Power comparisons of the seven tests for the power as a function of the percentage of protective variants and for the power as a function of the percentage of causal variants are given in Figures 2 and 3, respectively. These two figures show that the powers of PE, PE-TOW, SKAT, SKAT-C, and TOW are robust to the percentage of protective variants and the percentage of causal variants while powers of PE-WS and WS decrease with the increasing of the percentage of protective variants and increase with the increasing of the percentage of causal variants. Other patterns of power comparisons are similar to that in

Figure 1. We also provide power comparisons of the seven tests using a small significance level of 10^{-6} (Figure 4) and using a large sample size of 5000 (Figure 5). Figure 4 shows that the patterns of the power comparisons using significance level 10^{-6} are similar to that using a significance level of 0.05 in Figure 1 (Both; Without covariates). Figure 5 shows that the patterns of the power comparisons using a sample size of 5000 are similar to that using a sample size of 1000 in Figure 1 (Both; Without covariates).

In summary, PE-WS and PE-TOW are consistently more powerful than WS and TOW, respectively. When causal variants contain both common and rare variants, PE is the most powerful test, SKAT-C is more powerful than SKAT, and SKAT is more powerful than TOW. When causal variants are all rare variants, TOW is more powerful than SKAT, and SKAT is more powerful than SKAT-C. The powers of PE, PE-TOW, SKAT, SKAT-C, and TOW are robust to the percentage of protective variants and the percentage of causal variants.

1.5 Analysis of the GAW17 simulated dataset

The GAW17 simulated dataset consists of a collection of 697 unrelated individuals, their real genotypes and 200 replicates of the simulated phenotypes. Three quantitative traits Q1, Q2 and Q4 are simulated. Covariates include age, sex, and smoking status. Since quantitative trait Q4 has no genetic components, we do not consider Q4 for the purpose of power comparisons. We perform power comparisons of the seven tests using quantitative traits Q1 and Q2. The p-values of TOW, PE-TOW, PE-WS, and PE are evaluated by 10,000 permutations and the p-values of WS, SKAT-C, and SKAT are evaluated by asymptotic distributions. The powers of the seven tests are calculated at a significance level of 0.001. We merge every two replicates to one replicate to increase the sample size. In all cases, the minor allele is associated with higher means of the two quantitative traits, which means that all causal variants are risk variants. Q1 has 9 causal genes and Q2 has 13 causal genes. We omit causal genes that have one variant, causal genes in which all of the seven tests have 100% power, and causal genes in which all of the seven tests have a power less than 10%. Q1 has 5 causal genes left and Q2 has 7 causal genes left. The powers of TOW, WS, and SKAT to test the association between each of the five causal genes and Q1 are not consistent with that in Table 2 of Sha *et al*. [2012] because we found that Sha *et al*. [2012] did not adjust trait values and genotypes for covariates when testing the association for Q1.

The powers of the seven tests to detect association between each of the 12 causal genes and Q1 or Q2 are given in Table 2. As shown in Table 2, WS, TOW, or SKAT-C is the

most powerful test in 1 out of 12 genes, PE-WS, PE-TOW, or SKAT is the most powerful test in 2 out of 12 genes, and PE is the most powerful test in 4 out of 12 genes. 3 out of 4 causal genes in which PE or SKAT-C is the most powerful test include common causal variants. Causal variants in the 6 genes in which TOW, WS, PE-TOW, or PE-WS is the most powerful test are all rare variants (MAF < 1%). Each of the 2 genes in which SKAT is the most powerful test contains causal variants with MAF in (0.01,0.035). Comparing TOW and WS with PE-TOW and PE-WS, PE-TOW is more powerful than TOW and PE-WS is more powerful than WS in 9 out of 12 causal genes. The results from the analysis of the GAW17 simulated dataset are consistent with those from the simulation studies.

1.6 Discussion

Based on cross-validation prediction error under Ridge regression, we developed novel statistical tests to test the association between variants (both common and rare variants) in a genomic region and a complex trait of interest. We proposed PE method based on Ridge regression. Combined with the weighting schemes, we further developed PE-WS and PE-TOW methods. We used extensive simulation studies to compare the performance of PE, PE-WS and PE-TOW with that of the existing methods: SKAT, SKAT-C, WS, and TOW. Our results showed that (1) the prediction error versions of TOW and WS (PE-TOW and PE-WS) are consistently more powerful than TOW and WS, respectively; (2) when causal variants contain both common and rare variants, PE is the most powerful test, SKAT-C is more powerful than SKAT, and SKAT is more powerful than TOW. When causal variants are all rare variants, TOW is more powerful than SKAT, and SKAT is more powerful than SKAT-C; and (3) the powers of PE, PE-TOW, SKAT, SKAT-C, and TOW are robust to the percentage of protective variants and the percentage of causal variants.

Each of the three proposed methods PE, PE-TOW, and PE-WS has its advantages in some scenarios. PE is more powerful than PE-TOW and PE-WS when causal variants contain both common and rare variants. PE-WS is a burden test and is more powerful than PE-TOW when the percentage of causal variants is large and the directions of the effects of the causal variants are all the same. PE-TOW is more powerful than PE-WS when the percentage of causal variants is small or the directions of the effects of the causal variants

are different. We may construct a robust test aiming to have the advantages of all of PE, PE-TOW and PE-WS. Let p_{PE}, p_{PE-TOW}, and p_{PE-WS} denote the *P*-values of PE, PE-TOW and PE-WS, respectively. Then, we define the test statistic of the robust test as $T_{robust} = \min\{p_{PE}, p_{PE-TOW}, p_{PE-WS}\}$. However, the performance of the robust test needs further investigation.

PE test statistic does not work well for rare variants. The reason is that many rare variants are singletons. From our simulation results, PE method may be more powerful than existing methods for common variants. The performance of PE for common variants needs further investigation.

Among the three proposed tests (PE, PE-WS, and PE-TOW), PE is most computationally intensive. The computation time required for running PE depends on the sample size, the number of variants in the genomic region, and the number of permutations. The running time of PE with 1000 permutations on a data set with 1000 individuals and 100 variants in a genomic region on a laptop with 4 Intel Cores @ 3.30GHz and 4 GB memory is about 0.1s. To perform genome-wide association studies, we can first select genomic regions that show evidence of association based on a small number of permutations (e.g. 1,000), and then a large number of permutations are used to test the selected regions.

1.7 Tables and Figures

Table 1.1. Type I error rates of the three proposed methods with 10,000 replicates

Significance level	Sample size	With covariates			Without covariates		
		PE-WS	PE-TOW	PE	PE-WS	PE-TOW	PE
0.05	500	0.0545	0.0485	0.0525	0.049	0.0506	0.0504
	1000	0.0503	0.051	0.0519	0.0493	0.0517	0.05
0.01	500	0.0104	0.0091	0.0107	0.0099	0.0088	0.0103
	1000	0.0112	0.010	0.0102	0.009	0.0097	0.0103
0.001	500	0.0007	0.0009	0.0006	0.0008	0.001	0.0011
	1000	0.0017	0.0005	0.0016	0.0011	0.0009	0.0008

Table 1.2. Power of the seven tests to detect the association between each of the five causal genes and quantitative trait Q1 and between each of the seven causal genes and quantitative trait Q2.

Traits	Gene Name	No. of Variants, No. of Causal Variants	Min, Max, Mean MAF	WS	TOW	PE-WS	PE-TOW	PE	SKAT-C	SKAT
Q1	ARNT	18, 5	0.07, 1.15, 0.33	0.08	0.52	0.05	0.55	0.83	0.84	**0.95**
	ELAVL4	10, 2	0.07, 0.07, 0.07	0.44	0.40	0.48	0.53	**0.72**	0.26	0.00
	FLT4	10, 2	0.07, 0.14, 0.11	**0.85**	0.68	0.70	0.51	0.79	0.14	0.68
	HIF1A	8, 4	0.07, 1.22, 0.39	0.63	0.56	0.46	0.43	**0.91**	0.66	0.88
	VEGFA	6, 1	0.22, 0.22, 0.22	0.33	0.16	**0.45**	0.22	0.22	0.23	0.10
Q2	BCHE	29, 13	0.07, 0.29, 0.10	0.20	0.34	0.23	**0.38**	0.27	0.02	0.14
	LPL	20, 3	0.07, 1.58, 0.60	0.01	0.23	0.05	0.28	0.16	0.33	**0.41**
	PDGFD	11, 4	0.07, 0.86, 0.29	0.07	0.23	0.09	**0.25**	0.15	0.04	0.15
	SIRT1	24, 9	0.07, 0.22, 0.12	0.50	**0.65**	0.52	0.60	0.41	0.61	0.55
	SREBF1	24, 10	0.07, 0.43, 0.22	0.26	0.20	**0.27**	0.25	0.19	0.03	0.06
	VNN1	7, 2	0.57, 17.1, 8.82	0.06	0.68	0.08	0.78	**0.95**	**0.95**	0.02
	VNN3	15, 7	0.07, 9.83, 2.06	0.33	0.55	0.35	0.66	**0.73**	0.68	0.40

Note: Min, Max and Mean MAF represent the minimum, maximum, and mean MAF (in percentage) at the causal variants. In each row, the boldfaced number represents the highest power in the row.

Figure 1.1. Power comparisons of the seven tests (PE-WS, PE-TOW, PE, WS, TOW, SKAT and SKAT-C) for the power as a function of heritability. “Rare” means that all causal variants are rare. “Both” means that causal variants contain both rare and common (1 common variant) and the heritability of the common variant is as twice as the heritability of all the rare causal variants. x axis represents the total heritability of all causal variants. Sample size is 1000. In this set of simulations, all causal variants are risk variants and 20% of rare variants are causal. The powers are evaluated at a significance level of 0.05.

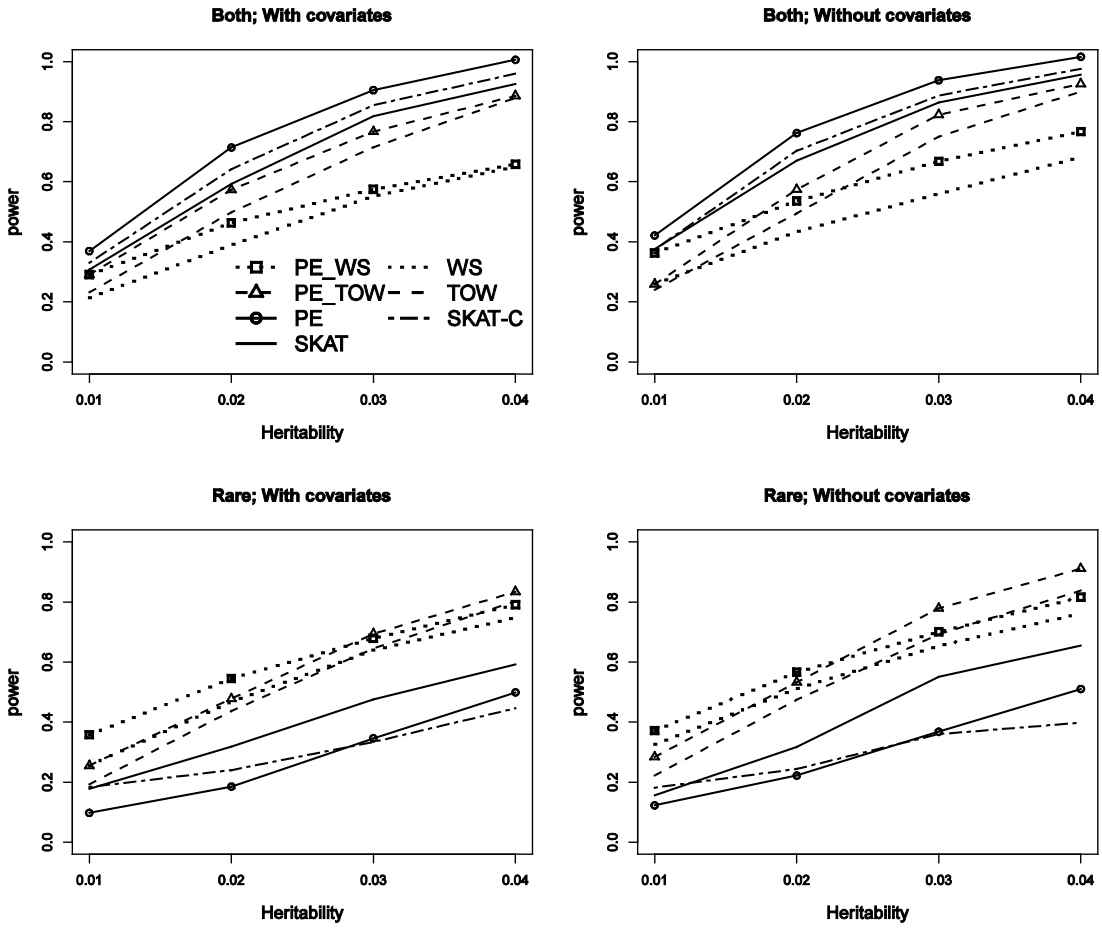


Figure 1.2. Power comparisons of the seven tests (PE-WS, PE-TOW, PE, WS, TOW, SKAT and SKAT-C) for the power as a function of the percentage of protective variants. "Rare" means that all causal variants are rare. "Both" means that causal variants contain both rare and common (1 common variant) and the heritability of the common variant is as twice as the heritability of all the rare causal variants. x axis represents the percentage of protective variants. Sample size is 1000. The total heritability is 0.03. 20% of rare variants are causal. The powers are evaluated at a significance level of 0.05.

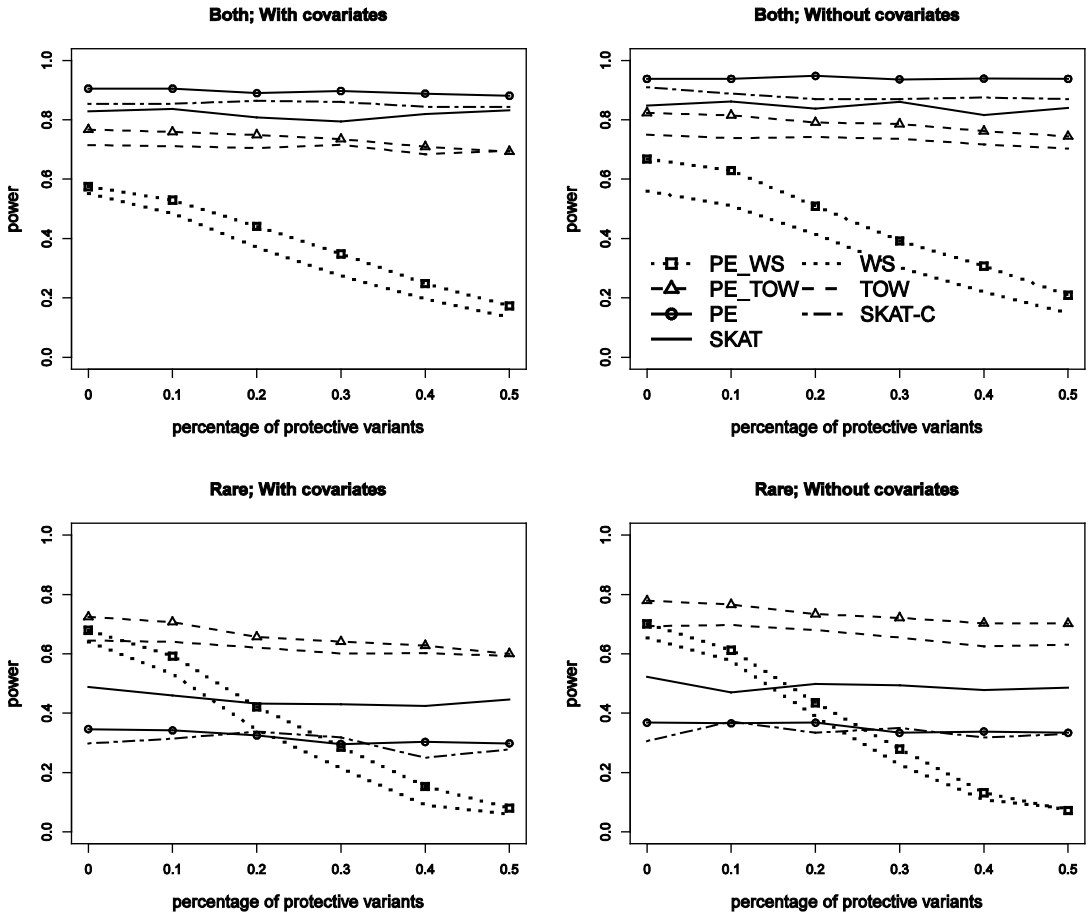


Figure 1.3. Power comparisons of the seven tests (PE-WS, PE-TOW, PE, WS, TOW, SKAT and SKAT-C) for the power as a function of the percentage of causal variants. "Rare" means that all causal variants are rare. "Both" means that causal variants contain both rare and common (1 common variant) and the heritability of the common variant is as twice as the heritability of all the rare causal variants. x axis represents the total heritability of all causal variants. Sample size is 1000. The total heritability is 0.03. All causal variants are risk variants. The powers are evaluated at a significance level of 0.05.

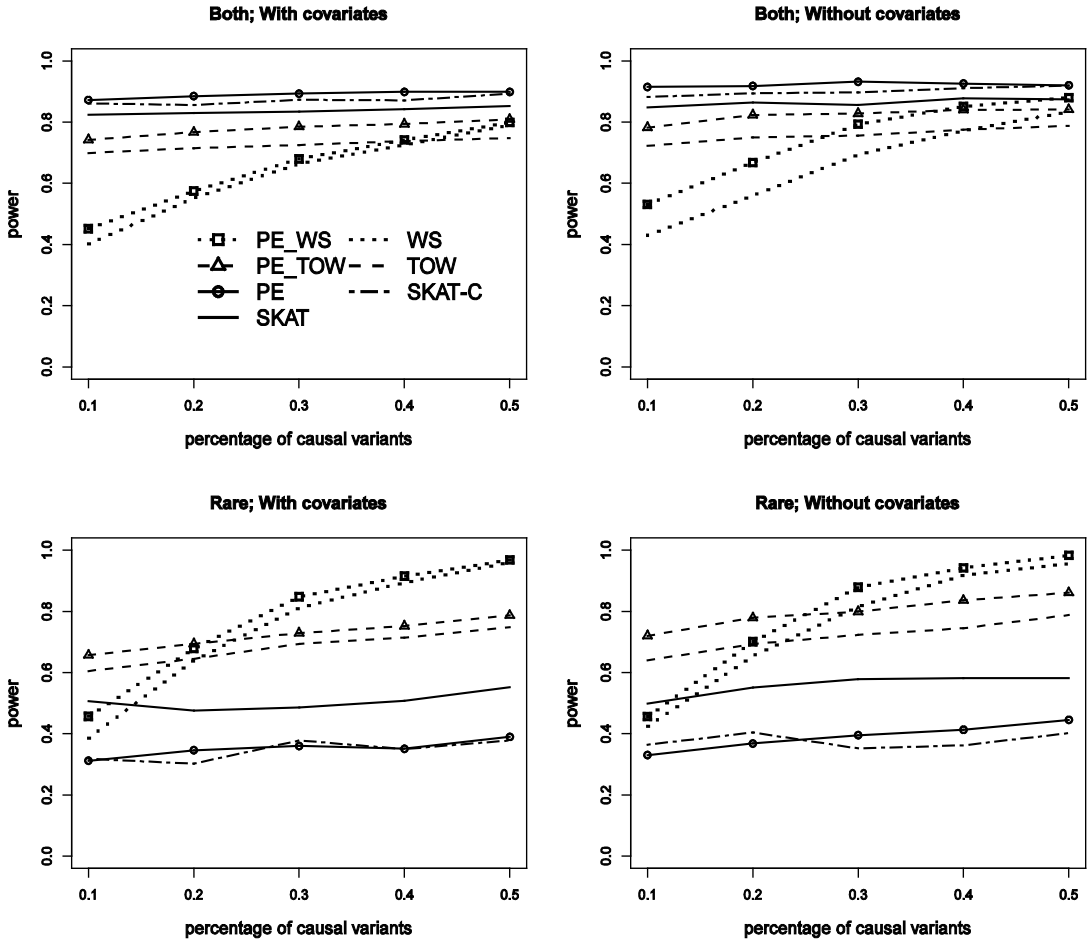


Figure 1.4. Power comparisons of the seven tests (PE-WS, PE-TOW, PE, WS, TOW, SKAT and SKAT-C) for the power as a function of heritability. "Both" means that causal variants contain both rare and common (1 common variant) and the heritability of the common variant is as twice as the heritability of all the rare causal variants. x axis represents the total heritability of all causal variants. Sample size is 1000. In this set of simulations, all causal variants are risk variants and 20% of rare variants are causal. Powers are evaluated at significance level 10^{-6} and p-values of PE-WS, PE-TOW, PE and TOW are evaluated by 10^7 permutations.

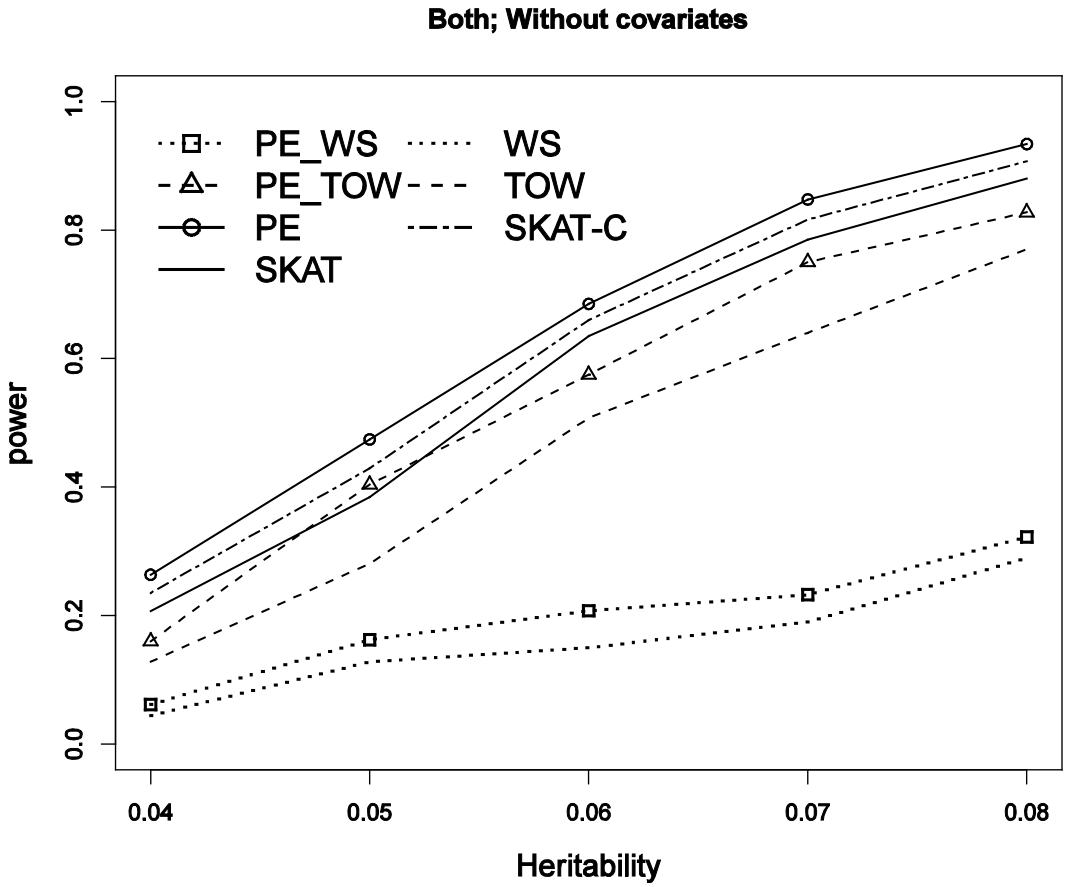
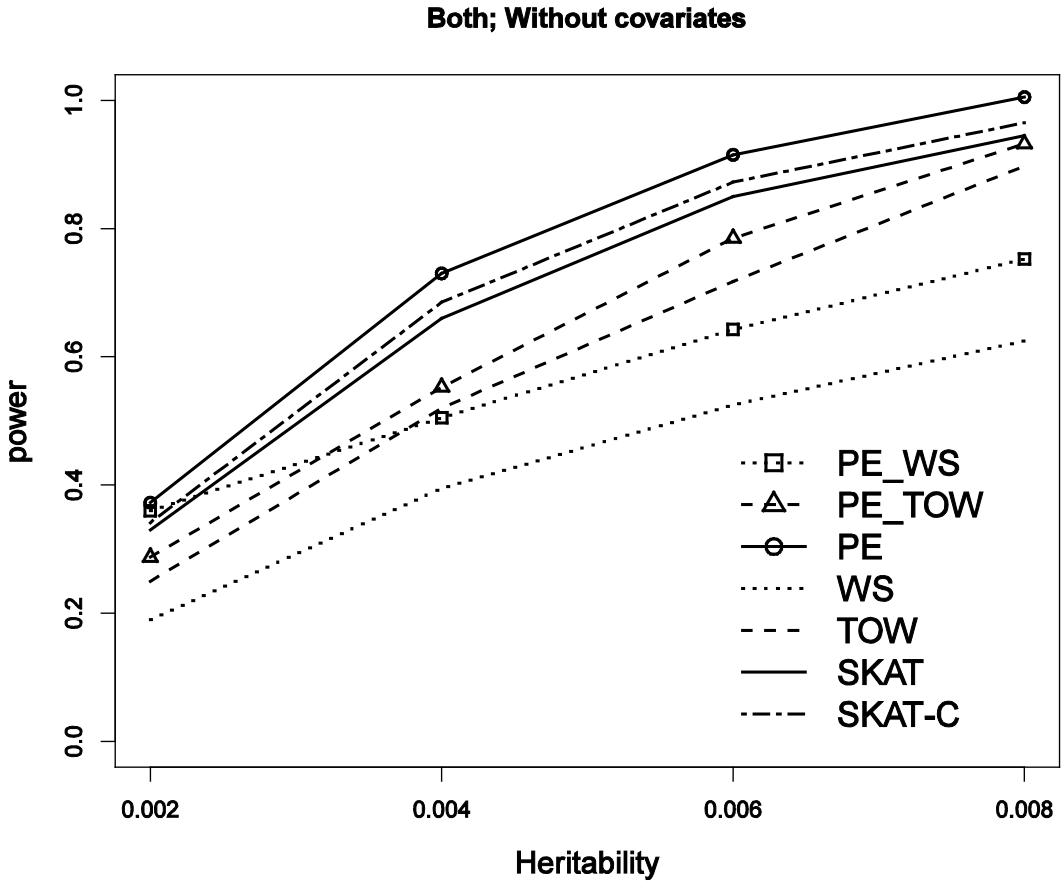


Figure 1.5. Power comparisons of the seven tests (PE-WS, PE-TOW, PE, WS, TOW, SKAT and SKAT-C) for the power as a function of heritability. “Both” means that causal variants contain both rare and common (1 common variant) and the heritability of the common variant is as twice as the heritability of all the rare causal variants. x axis represents the total heritability of all causal variants. Sample size is 5000. In this set of simulations, all causal variants are risk variants and 20% of rare variants are causal. The powers are evaluated at a significance level of 0.05.



2 Chapter 2

Joint Analysis of Multiple Phenotypes in Association Studies based on Cross-Validation Prediction Error

In genome-wide association studies (GWAS), the joint analysis of multiple phenotypes could have increased power over analyzing each phenotype individually. With this motivation, several methods that jointly analyze multiple phenotypes have been developed, such as O'Brien's method, Trait-based Association Test that uses Extended Simes procedure (TATES), MAONVA and MultiPhen. However, the performance of these methods under a wide range of scenarios is not consistent: one test may be powerful in some situations, but not in the others. Thus, one challenge in joint analysis of multiple phenotypes is to construct a test that could maintain good performance across different scenarios. In this article, we propose a novel statistical method to test the association between a genetic variant and multiple phenotypes based on cross-validation prediction error (PE). Extensive simulations are conducted to evaluate the type I error rates and to compare the power performance of the PE method with various existing methods. We show that the PE method controls the type I error rates very well and has consistently higher power than the tests we compared in all the scenarios. We conclude with the recommendation for the use of the PE method for its high and complementary performance.

2.1 Introduction

Traditionally, genome-wide association studies (GWAS) have performed on individual phenotype. In spite of the success of GWAS in identifying thousands of associations between single nucleotide polymorphism (SNPs) and complex diseases, these identified variants only contribute a small portion of the phenotypic variance. In the study of a complex disease, several correlated phenotypes are usually measured for a disorder or its risk factors [Yang et al., 2010]. Jointly using multiple correlated phenotypes can help to increase statistical power to detect causal variants and illuminate on underlying biological mechanisms.

One method to use multiple phenotypes in association study is to perform each phenotype separately as standard univariate association test and then aggregate the results, while this approach will have a loss in power due to the penalties from the multiple testing [O'Reilly et al. 2012; Yang, 2010] and the ignorance of the correlation structure among phenotypes [Wang et al. 2015, Yang et al. 2016]. Thus, multiple-phenotype association study that uses multiple phenotypes simultaneously has become popular.

Several methods to detect association using multiple phenotypes have been introduced in recent years. For example, O'Brien method (OB) is proposed to combines test statistics obtained from association test for each individual phenotype [O'Brien et al. 1984]. OB is the most powerful test when the genetic effects are homogeneous, however, this method lose power when genetic effects are heterogeneous, especially when genetic effects have

opposite directions [Yang et al. 2010; Zhu et al. 2016]. The canonical correlation analysis (CCA) proposed by Ferreira & Purcell (2008), conducts the linear combination of phenotypes that explain the largest possible amount of the covariation between a genetic variant and phenotypes [Ferreira et al. 2008]. We could also use multivariate analysis of variance (MANOVA) in regression to study multiple phenotypes [Cole et al., 1994]. MANOVA is equivalent to canonical correlation analysis when canonical correlation analysis is applied to a single variant [Galesloot et al. 2014]. MultiPhen, which proposed by O'Reilly et al. [2012], can be used to detect the association between one SNP and multiple phenotypes by reversing response and predictors via a proportional odds regression model. And when a small number of phenotypes are included, MultiPhen and MANOVA lead to similar performance [Aschard et al. 2014, Zhu et al. 2016]. MANOVA and CCA require the assumption of normality of multiple phenotypes, while MultiPhen has no inflated type I error rates on non-normal phenotypes. van der Sluis et al. [2013] proposed a trait-based association test using an extended Simes procedure (TATES) that conducts association test for each phenotype and then combines the univariate p-values with corrected correlations among phenotypes [van der Sluis et al. 2013]. Some other variable reduction methods have been proposed to test for the association between a genetic variant and the linear combination of the multiple phenotypes rather than the original phenotypes [Zhu et al., 2018, Wang et al. 2008, Klei et al. 2008]. For example, principal component of phenotypes (PCP) that maximizes the phenotype variation, is the most popular dimension reduction method [Wang et al. 2008]. Based on PCP, Klei et al.

[2008] developed principal component of heritability (PCH) by maximizing the heritability among all linear combination of the phenotypes.

Although there are many proposed methods for joint analysis of multiple phenotypes, the performance of these methods under a wide range of scenarios is not consistent [Zhu et al. 2016]: one test may be powerful in some situations, but not in the others. Thus, one challenge in multiple phenotype analysis is to construct a test that could maintain good performance across different scenarios. In this article, we develop a novel statistical method to test the association between a genetic variant and multiple phenotypes based on cross-validation prediction error (PE). Extensive simulation studies are conducted to evaluate the type I error rates and to compare power performance of the PE method with various existing methods. We show that the PE method controls the type I error rates very well and has consistently higher power than other methods we compared in all the scenarios.

2.2 Method

Prediction Error Model

We consider a sample with n unrelated individuals. Each individual has K (potentially correlated) phenotypes and has been genotyped at a variant of interest. Let y_{ik} denote the k^{th} phenotype value of the i^{th} individual and x_i denote the genotype score of the i^{th} individual, where $x_i \in \{0,1,2\}$ is the number of minor alleles that the i^{th} individual carries. We model the relationship between the multiple phenotypes and the variant using the inverse linear regression model

$$x_i = \beta_0 + \beta_1 y_{i1} + \cdots + \beta_K y_{iK} + \varepsilon_i . \quad (1)$$

Let $y_i = \left(1, y_{i1}, \ldots, y_{iK}\right)^T$, $\beta = (\beta_0, \beta_1, \ldots, \beta_K)^T$ and $x = \left(x_1, \ldots, x_n\right)^T$. To test the association between the K multiple phenotypes and the variant, we test the null hypothesis $H_0 : \beta_1 = \cdots = \beta_K = 0$ under equation (1).

When we conduct analysis on the model, the parameter estimation make the model fit a particular dataset as well as possible, it might cause some problems like overfitting. In order to have an insight on how generally the model will fit to an independent data and how reliably the model will predict future observations, we use cross validation to improve the model performance. We apply k-fold cross-validation, divide the data into k equal parts, and use each of the k parts as the testing set and other k-1 parts as the training

set. We use the training set to estimate β , and use the prediction equation $\hat{x}_i = y_i^T \hat{\beta}$ to predict genotypes in the testing set.

In the k-fold cross-validation, leave-one-out cross-validation (LOOCV) (k is equal to n) is an extreme case. LOOCV takes advantage of almost the entire data set, the difference between the training set in each fold and the entire dataset is only one single pattern. Therefore, LOOCV is approximately unbiased [Luntz et al. 1969]. Also, since each sample is used for both training and testing, there is no randomness of allotting some samples for training and others for testing, which leads to overall less variability than k-fold cross-validation method (when $k < n$). Furthermore, the LOOCV has a fast algorithm to calculate the cross-validation prediction error [James et al. 2013]. In conclusion, our proposed test is based on LOOCV.

We construct a novel statistical test to test the association between the multiple phenotypes and the variant in a genomic region based on LOOCV prediction error. We propose to use the LOOCV prediction error under model (1) as a test statistic. Let $\hat{x}_{-i}$ denote the LOOCV predicted value (leave the i^{th} individual out) of x_i under model (1). Then, the statistic can be written as

$$T = \sum_{i=1}^{n} \left(x_i - \hat{x}_{-i} \right)^2 \tag{2}$$

Note that low values of T would imply significance.

Parameter estimation

When the multiple phenotypes are highly correlated with each other, the rank of matrix $Y=\left(y_1,\ldots,y_n\right)^T$ will be less than K, then the inverse of Y^TY doesn't exist, which results in that the ordinary linear square estimate may not be unique. We can use penalized regression, such as Ridge regression [Halawa & Bassiouni, 2000; Hoerl, Kannard, & Baldwin, 1975] and Lasso regression [Meier, Van De Geer, & Bühlmann, 2008; Tibshirani, 1996; Yuan & Lin, 2006] to make sure that we can find a solution. For example, in Ridge regression, we are introducing a diagonal matrix λI along the matrix Y^TY, which guarantees the matrix $Y^TY+\lambda I$ to be invertible. Penalized regressions have been applied to the analysis of genetic data [Ayers & Cordell, 2013, 2010; Cule & De Iorio, 2013; Cule, Vineis, & De Iorio, 2011; Malo, Libiger, & Schork, 2008; Warren, Casas, Hingorani, Dudbridge, & Whittaker, 2014, Yang et al., 2017]. In this paper, we propose to use Ridge regression, because another benefit of Ridge regression is reducing overfitting through penalizing the size of the regression coefficients and yielding better predictions than ordinary linear square estimation via a tradeoff between bias and variance.

In the regression model $x_i=y_i^T\beta+\varepsilon_i$, the ridge regression estimator $\hat{\beta}$ is defined as the value of β that minimizes $\sum_i(x_i-y_i^T\beta)^2+\lambda\sum_j\beta_j^2$. The solution to the ridge regression problem is given by $\hat{\beta}_\lambda=(Y^TY+\lambda I)^{-1}Y^Tx$, where $Y=\left(y_1,\ldots,y_n\right)^T$, and $\lambda\geq 0$ is a tuning

parameter. For Ridge regression, let $\hat{x}_{-i}^{\lambda}$ denote the LOOCV predicted value (leave the i^{th} individual out) of x_i under model (1). Then, the statistic can be written as $T_\lambda = \sum_{i=1}^{n}\left(x_i - \hat{x}_{-i}^{\lambda}\right)^2$ .We denote the p-value of T_λ as p_λ . We define the LOOCV Prediction Error test statistic (PE) as

$$T_{PE} = \min_\lambda p_\lambda . \tag{3}$$

In this study, we use a simple grid search method to evaluate the minimization. We divide the interval $[0,\infty)$ into subintervals $0 \le \lambda_1 < ... < \lambda_{L-1} < \lambda_L < \infty$. In the simulation studies, we used $L = 8$ and $\left(\log\lambda_1,\ldots,\log\lambda_8\right) = \left(0,1,2,3,3.5,4,4.5,5\right)$. Then, $T_{PE} = \min_\lambda p_\lambda = \min_{1\le l\le L} p_{\lambda_l}$.

We use the same permutation procedure to evaluate the p-value of T_{PE} as in Yang *et al*. [2017]. Intuitively, two layers of permutations are needed to estimate p_{λ_l} and the overall p-value for the test statistic T_{PE} . Actually, we can use one layer of permutation to estimate p_{λ_k} and the overall p-value for the test statistic T_{PE} [Ge *et al*., 2003; Yang et al. 2017]. Suppose that we perform permutation B times. In each permutation, we randomly shuffle the individual genotypes. Let $T_{\lambda_l}^{(b)}$ denote the values of T_{λ_l} based on the b^{th} permuted data for $b = 0,1,...,B$ and $l = 1,\ldots,L$, where $b = 0$ represents the original data. Then, we transfer $T_{\lambda_l}^{(b)}$ to $p_{\lambda_l}^{(b)}$ by

$$p_{\lambda_l}^{(b)} = \frac{\#\{d : T_{\lambda_l}^{(d)} < T_{\lambda_k}^{(b)} \text{ for } d = 1,\ldots,B\}}{f(b)} \tag{4}$$

where $f(0) = B$ and $f(b) = B - 1$ for $b = 1,\ldots,B$. Let $p^{(b)} = \min_{1 \le l \le L} p_{\lambda_l}^{(b)}$. Then, the p-value of T_{PE} is given by

$$\frac{\#\{b : p^{(b)} < p^{(0)} \text{ for } b = 1,2,\ldots,B\}}{B}.$$

In the next section, we propose an algorithm that can perform the permutation procedure described above more efficiently.

2.3 A fast algorithm for the permutation procedure

Let $y_i = (1, y_{i1}, \ldots y_{iK})^T$, $Y = (y_1, \ldots, y_n)^T$, $A_\lambda = (Y^T Y + \lambda I)^{-1}$, $h_i^\lambda = y_i^T A_\lambda y_i$, $h_\lambda = (h_1^\lambda, ..., h_n^\lambda)$, and $\hat{\beta}_\lambda = A_\lambda Y^T x$. Then, the Ridge predicted values are $\hat{x}_i^\lambda = y_i^T \hat{\beta}_\lambda$ and $\hat{x}_\lambda = (\hat{x}_1^\lambda, \ldots, \hat{x}_n^\lambda)^T = Y(Y^T Y + \lambda I)^{-1} Y^T x$. We can show that LOOCV prediction error in Ridge regression has a closed-form formula, that is, $x_i - \hat{x}_{-i}^\lambda = (x_i - \hat{x}_i^\lambda) / (1 - h_i^\lambda)$. For two matrices or vectors A and B, we use $A * B$ and $\frac{A}{B}$ to denote the element-wise operations. We assume $n \geq K+1$. We perform singular value decomposition of Y, that is, $Y = UDV$, where U is an $n \times (K+1)$ matrix with orthonormal columns, D is $(K+1) \times (K+1)$ diagonal matrix with non-negative real numbers on the diagonal, and V is an $(K+1) \times (K+1)$ orthogonal matrix. Let $D = diag(d_1, \ldots, d_{K+1})$. Then, $\hat{x}_\lambda = UC_\lambda U^T x$, where $C_\lambda = diag(c_{\lambda,1}, \ldots, c_{\lambda,K+1})$ and $c_{\lambda,j} = d_j^2 / (d_j^2 + \lambda)$ for $j = 1, \ldots, K+1$. Let $c_\lambda = (c_{\lambda,1}, \ldots, c_{\lambda,K+1})^T$ and $x^{(K)} = U^T x$ be a $K+1$ dimensional vector. Then, $\hat{x}_\lambda = UC_\lambda x^{(K)} = U(c_\lambda * x^{(K)})$ and $h_\lambda = diag(UC_\lambda U^T)$. For $0 \leq \lambda_1 < \ldots < \lambda_L < \infty$, let $C = (c_{\lambda_1}, \ldots, c_{\lambda_L})$ and $H = (h_{\lambda_1}, \ldots, h_{\lambda_L})$. Then, $(\hat{x}_{\lambda_1}, \ldots, \hat{x}_{\lambda_L}) = U(C * x^{(K)}) = U(c_{\lambda_1} * x^{(K)}, \ldots, c_{\lambda_L} * x^{(K)})$. If we denote $B = \frac{x - \hat{X}}{1 - H} = \frac{(x - \hat{x}_{\lambda_1}, \ldots, x - \hat{x}_{\lambda_L})}{1 - H}$, then $(T_{\lambda_1}, \ldots, T_{\lambda_L}) = colSums(B * B)$. Note that C, U, and H only depend on phenotypes and $\lambda_1, \ldots, \lambda_L$. Thus, C, U, and H do not change in each permutation. For a GWAS, C, U, and H also do not change at different SNPs. To

perform GWASs, we can first select SNPs that show evidence of association based on a small number of permutations (e.g. 1,000), and then a large number of permutations are used to test the selected SNPs.

With the fast algorithm, we can use less than one day to perform a typical GWAS. Our preliminary studies showed that performing a GWAS with 5,430 individuals and 7 phenotypes only needs 10 hours on Intel Xeon E5-2680v3 by using a single node.

2.4 Comparison of Methods

We compare the performance of the proposed test, PE, with that of the O'Brien's method (OB) [O'Brien et al. 1984], Trait-based Association Test that uses Extended Simes procedure (TATES) [van der Sluis et al. 2013], Optimal weight method (OW) [Zhu et al., 2016], Multivariate analysis of variance (MANOVA) [Cole et al., 1994], and Joint model of multiple phenotypes (MultiPhen)[O'Reilly et al.2012].

2.5 Simulation Study

In simulation studies, the simulation settings are similar to that of Wang *et al.* (2016). We evaluate Type I error rates of PE method by generating unrelated data sets with three different sample sizes, 500, 1,000 and 2,000. For power comparison, we compare powers of different methods by simulation data sets with 1,000 unrelated individuals.

For the genotype data, we generate genotype at one variant by assuming Hardy-Weinberg Equilibrium and according to minor allele frequency (MAF). For each individual, we generate K phenotypes (K = 20 or 40). The K phenotypes are generated from the following model

$$y = \phi x + c\boldsymbol{\gamma}\omega + \sqrt{1-c^2} \times \varepsilon \tag{5}$$

where $y = (y_1, ..., y_K)^T$. $\phi = (\phi_1, ..., \phi_K)$ are the genetic effects of a variant of interest on the K phenotypes; x is the genotypic score at the variant; c is a constant number; γ is a $K \times R$ matrix; $\omega = (\omega_1, ..., \omega_R)^T$ is a vector of factors with R elements and $\omega = (\omega_1, ..., \omega_R)^T \sim MVN(0, \Sigma)$, $\Sigma = \rho A + (1-\rho)I$, ρ is the correlation between factors, A is a matrix with elements of 1, and I is the identity matrix; $\varepsilon = (\varepsilon_1, ..., \varepsilon_K)^T$ is a vector of residuals, $\varepsilon_1, ..., \varepsilon_K$ are independent, and $\varepsilon_k \sim N(0,1)$ for $k = 1, ..., K$. Based on equation (4), we consider the following four models in which the within-factor correlation is c^2 and the between-factor correlation is ρc^2.

Model 1: There is only one factor and genotypes impact on all phenotypes with a different effect size. That is, $R=1$, $\omega=\beta(1,2,...,K)^T$, and $\boldsymbol{\gamma}=(1,...,1)^T$.

Model 2: There are two factors and genotypes impact on one factor. That is, $R=2$, $\lambda=\left(0,\ldots,0,\underbrace{\beta,\ldots,\beta}_{K/2}\right)^T$, and $\boldsymbol{\gamma}=diag\left(D_1,D_2\right)$, where $D_i=\left(\underbrace{1,\ldots,1}_{K/2}\right)^T$ for $i=1,2$.

Model 3: There are five factors and genotypes impact on two factors. That is, $R=5$, $\lambda=\left(\beta_{11},\ldots,\beta_{1k},\beta_{21},\ldots,\beta_{2k},\beta_{31},\ldots,\beta_{3k},\beta_{41},\ldots,\beta_{4k},\beta_{51},\ldots,\beta_{5k}\right)^T$, and $\boldsymbol{\gamma}=diag\left(D_1,D_2,D_3,D_4,D_5\right)$, where $D_i=\left(\underbrace{1,\ldots,1}_{K/5}\right)^T$ for $i=1,\ldots,5$; $k=K/5$; $\beta_{11}=\cdots=\beta_{1k}=\beta_{21}=\cdots=\beta_{2k}=\beta_{31}=\cdots=\beta_{3k}=0$; $\beta_{41}=\cdots=\beta_{4k}=-\beta$; and $\left(\beta_{51},\ldots,\beta_{5k}\right)=\frac{2\beta}{k+1}\left(1,\ldots,k\right)$.

Model 4: There are five factors and genotypes impact on four factors. That is, $R=5$, $\lambda=\left(\beta_{11},\ldots,\beta_{1k},\beta_{21},\ldots,\beta_{2k},\beta_{31},\ldots,\beta_{3k},\beta_{41},\ldots,\beta_{4k},\beta_{51},\ldots,\beta_{5k}\right)^T$, and $\boldsymbol{\gamma}=diag\left(D_1,D_2,D_3,D_4,D_5\right)$, where $D_i=\left(\underbrace{1,\ldots,1}_{K/5}\right)^T$ for $i=1,\ldots,5$; $k=K/5$; $\beta_{11}=\cdots=\beta_{1k}=0$; $\beta_{21}=\cdots=\beta_{2k}=\beta$; $\beta_{31}=\cdots=\beta_{3k}=-\beta$; $\left(\beta_{41},\ldots,\beta_{4k}\right)=-\frac{2\beta}{k+1}\left(1,\ldots,k\right)$; and $\left(\beta_{51},\ldots,\beta_{5k}\right)=\frac{2\beta}{k+1}\left(1,\ldots,k\right)$.

For the Type I error rates, we set $\beta = 0$ to indicate that the genetic variant has no effect on all phenotypes. For power comparisons, we consider different values of β. To evaluate Type I error rate and power, we set MAF = 0.3, the between-factor correlation is 0.15, and the within-factor correlation is 0.25.

2.6 Simulation Result

To evaluate the Type I error of the proposed PE methods, we consider different significance levels (0.01 and 0.05), different sample sizes (500, 1000 and 2000), and different number of phenotypes (10, 20 and 40). The p-values of PE are calculated using 1,000 permutations. Type I error rates are evaluated using 10,000 replicated samples. For 10,000 replicated samples, the 95% confidence intervals (CIs) for the estimated type I error rates of nominal levels 0.05 and 0.01 are (0.04562, 0.05438) and (0.00804, 0.01196), respectively. The estimated type I error rates of the proposed test are summarized in Table 1. From this table, we can see that all of the estimated type I error rates are within 95% CIs which indicates that the proposed method is valid.

In power comparisons, the p-values of PE are calculated using 1,000 permutations, the p-values of MultiPhen, OW, TATES, MANOVA, OB are evaluated using asymptotic distribution. The powers of all of the six tests are evaluated using 1,000 replicated samples at a significance level of 0.05.

Figures 1 and 2 show the power comparisons of the six methods as a function of effect size β with $K = 20$ and 40, respectively. As shown in two figures,

(1) PE is the most powerful method, and the power of PE is much higher than the second powerful test.

(2) as effect size β increases, the powers of all tests increase as well; as K increases from 20 to 40, PE presents more ascendancy than the other tests.

(3) MultiPhen, OW and MANOVA have similar powers under all four models. A similar conclusion has been mentioned in some published papers [Zhu et al., 2016, van der Sluis et al. 2013, O'Reilly et al.2012].

(4) OB is comparable to MultiPhen, OW and MANOVA in model 1 and 2, but has almost no power when genetic effects are heterogeneous (model 3 and 4).

(5) when the genetic variant affects a portion of phenotypes (model 2), TATES is more powerful. It is because TATES only depends the strongest associated phenotypes.

Power comparisons of the six methods as a function of within factor correlation with $K = 20$ and 40 are given in Figure 3 and Figure 4, respectively. As shown in these two figures,

(1) PE is the most powerful test.

(2) the performance of TATES is relatively robust to the within factor correlation, especially when the effects have opposite directions.

(3) MultiPhen, OW and MANOVA have similar powers under all models. When the within factor correlation is increasing, the powers of these three tests have increasing trend or decreasing trend depending on the different model settings. This pattern has been confirmed in Zhu's paper [Zhu et al., 2016].

(4) OB is the least powerful test except under model 2 with the within factor correlation >0.2.

In summary, PE is consistently the most powerful test among the tests we compared under all simulation scenarios.

2.7 Real Data Analysis

Chronic obstructive pulmonary disease (COPD) is a terminology to describe progressive life-threatening lung diseases that causes breathlessness and serious illness, including emphysema, chronic bronchitis, refractory asthma and some forms of bronchiectasis. A global prevalence of 251 million cases of COPD is reported in 2016 and it is estimated that COPD caused 3.17 million deaths in 2015 [WHO, 2017]. The COPDGene aims to find inherited or genetic factors that associated with COPD. The COPDGene dataset includes 10,192 participants, 3,408 of them are African-Americans(AA), and 6,784 of them are non-Hispanic Whites(NHW). We select seven quantitative COPD-related phenotypes (FEV1, Emphysema, Emphysema Distribution, Gas Trapping, Airway Wall Area, Exacerbation frequency, and Six-minute walk distance) and 4 covariates (BMI, Age, Pack-Years and Sex) [Liang et al., 2016] in the following data analysis.

We deleted individuals and genotypes with missing data. A set of 5,430 non-Hispanic Whites across 630, 860 SNPs is used after excluding missing data. Then we adjusted the phenotypes for the covariates by apply a linear regression [Sha et al., 2012; Zhu et al., 2018], we regress each phenotype on covariates, replace original phenotypes with the residuals of the regression, and apply each of the six tests to detect the association between the covariates-adjusted phenotypes (residuals) and each SNP.

We use genome-wide significance level 5×10^{-8} to identify SNPs significantly associated with the seven COPD-related phenotypes. There are total 14 SNPs identified

by at least one method (Table 2). All of the 14 SNPs had been reported to be associated with COPD by previous studies [Brehm, et al., 2011; Cho, et al., 2010; Cui, et al., 2014; Du, et al., 2016; Hancock, et al., 2010; Li, et al., 2011; Lutz, et al., 2015; Pillai, et al., 2009; Wilk, et al., 2009; Wilk, et al., 2012; Young, et al., 2010; Zhang, et al., 2011; Zhu, et al., 2014].

As shown in Table 2, MultiPhen identified 14 SNPs; OW, MANOVA and PE identified 13 SNPs; TATES identified 9 SNPs; and O'Brien method did not identify any SNP. In summary, the number of SNPs identified by PE is comparable to the largest number of SNPs identified by other tests and the COPD analysis results are consistent with our simulation results.

2.8 Discussion

For complex diseases in GWAS, the association between one SNP and each phenotype is usually weak. Analyzing multiple disease-related phenotypes could increase the statistical power to identify the association between genotypes and complex diseases. In this article, we developed a novel statistical method, PE, to test the association between a genetic variant and multiple phenotypes based on cross-validation prediction error, and showed that the PE method controls the type I error rates very well and has consistently high power among all the scenarios. Overall, PE is the most powerful test and has much higher power than the second powerful test; OW, MANOVA, and MultiPhen have very

similar performance; OB loses power dramatically when genetics effects are heterogeneous, especially when opposite effect directions occur; TATES is relatively robust to within factor correlation and more powerful when the genetic effect only works on a portion of phenotypes. In real data analysis, PE identified 13 out of 14 significant SNPs, which is comparable to MultiPhen (14 out of 14).

In COPDGene study, we incorporated covariates by regressing the phenotypes on covariates and considered the residuals as adjusted phenotypes [Sha et al., 2012; Zhu et al., 2018]. And we removed the observations with any missing information. For COPD dataset, the missing rate is about 20%. Note that deletion of missing data results in the reducing of sample size which leads to loss of power. In real data analysis, if the missing rate is more than 30%, we can use imputation methods [Ali et al., 2011] to impute missing data.

In summary, we believe that PE is a recommended approach since it provided robust good performance to test multiple phenotypes with a genetic variant under different scenarios.

2.9 Tables and Figures

Table 2.1. Estimated Type I error rates for the PE method under four models. The Type I error rates are evaluated using 10,000 replicated samples. *P*-values of PE are estimated by 1,000 permutations. α is the significance level. For 10,000 replicated samples, the 95% confidence intervals (CIs) for Type I error rates at nominal levels 0.05 and 0.01 are (0.04562, 0.05438) and (0.00804, 0.01196), respectively.

Sample Size	Number of Phenotypes	Significance Level	Model 1	Model 2	Model 3	Model 4
500	10	$\alpha = 0.01$	0.0103	0.0109	0.0112	0.0094
		$\alpha = 0.05$	0.0480	0.0512	0.0523	0.0532
	20	$\alpha = 0.01$	0.0116	0.0107	0.0114	0.0112
		$\alpha = 0.05$	0.0503	0.0499	0.0473	0.0515
	40	$\alpha = 0.01$	0.0112	0.0118	0.0121	0.0103
		$\alpha = 0.05$	0.0524	0.0515	0.0518	0.0541
1000	10	$\alpha = 0.01$	0.0503	0.0499	0.0473	0.0479
		$\alpha = 0.05$	0.0535	0.0532	0.0514	0.0492
	20	$\alpha = 0.01$	0.0101	0.0095	0.0112	0.0083
		$\alpha = 0.05$	0.0500	0.0501	0.0524	0.0469
	40	$\alpha = 0.01$	0.0094	0.0116	0.0117	0.0105
		$\alpha = 0.05$	0.0472	0.0512	0.0514	0.0508
2000	10	$\alpha = 0.01$	0.0111	0.0094	0.0118	0.0094
		$\alpha = 0.05$	0.0489	0.0491	0.0508	0.0465
	20	$\alpha = 0.01$	0.0113	0.0107	0.0098	0.0108
		$\alpha = 0.05$	0.0513	0.0491	0.0516	0.0523
	40	$\alpha = 0.01$	0.0099	0.0091	0.0107	0.0110
		$\alpha = 0.05$	0.0498	0.0480	0.0492	0.0476

Table 2.2 Significant SNPs and the corresponding p-values in the analysis of COPDGene. The p-values of PE are evaluated using 10^8 permutations The p-values of OB, TATES, OW, MANOVA, and MultiPhen are evaluated using asymptotic distributions. The graying out p-values indicate the p-values $> 5 \times 10^{-8}$.

Chr	Position	Variant identifier	OB	TATES	OW	MANOVA	MultiPhen	PE
4	145431497	rs1512282	0.46	7.09×10^{-13}	8.10×10^{-14}	6.52×10^{-14}	1.03×10^{-9}	0
4	145434744	rs1032297	0.49	0	1.11×10^{-16}	1.11×10^{-16}	7.69×10^{-14}	0
4	145474473	rs1489759	0.42	0	1.11×10^{-16}	6.68×10^{-17}	1.22×10^{-16}	1.00×10^{-8}
4	145485738	rs1980057	0.49	0	1.11×10^{-16}	7.12×10^{-17}	8.14×10^{-17}	1.00×10^{-8}
4	145485915	rs7655625	0.34	6.11×10^{-9}	1.87×10^{-9}	1.69×10^{-9}	9.13×10^{-17}	5.00×10^{-8}
15	78882925	rs16969968	0.96	5.40×10^{-8}	2.05×10^{-11}	1.77×10^{-11}	7.84×10^{-12}	0
15	78894339	rs1051730	0.99	3.13×10^{-8}	1.54×10^{-11}	1.32×10^{-11}	8.16×10^{-12}	0
15	78898723	rs12914385	0.99	2.76×10^{-8}	1.64×10^{-11}	1.41×10^{-11}	1.48×10^{-12}	0
15	78911181	rs8040868	0.99	5.53×10^{-10}	2.09×10^{-12}	1.76×10^{-12}	2.59×10^{-12}	0
15	78878541	rs951266	0.77	2.55×10^{-9}	3.24×10^{-12}	2.74×10^{-12}	1.02×10^{-11}	0
15	78806023	rs8034191	0.87	1.06×10^{-7}	2.42×10^{-10}	2.14×10^{-10}	7.74×10^{-11}	0
15	78851615	rs2036527	0.88	1.62×10^{-7}	4.47×10^{-10}	3.99×10^{-10}	1.77×10^{-10}	0
15	78826180	rs931794	0.91	1.23×10^{-7}	2.64×10^{-10}	2.35×10^{-10}	9.09×10^{-11}	0
15	78740964	rs2568494	0.27	2.93×10^{-5}	1.12×10^{-7}	1.05×10^{-7}	4.23×10^{-8}	1.50×10^{-7}

Figure 2.1 Power comparisons of the six methods as a function of effect size β. The total number of phenotypes is $K = 20$, sample size is 1000, MAF is 0.3, the between-factor correlation is 0.15, and the within-factor correlation is 0.25. Significance is assessed at the 5% level.

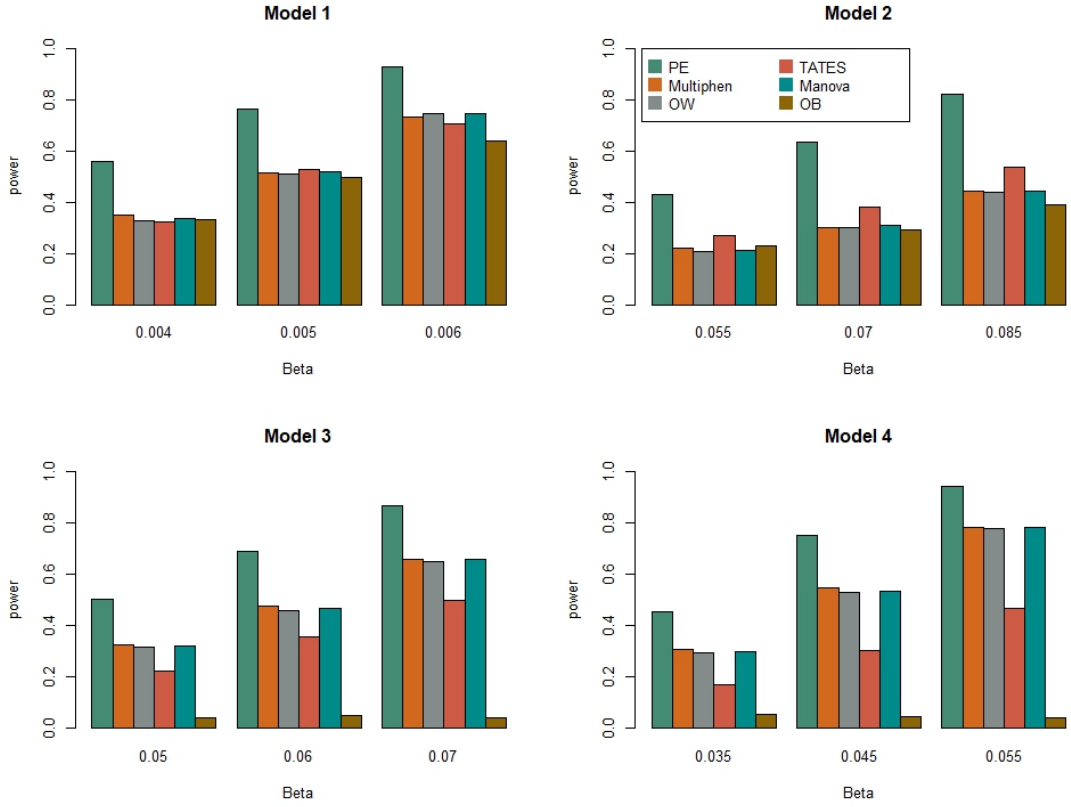


Figure 2.2. Power comparisons of the six methods as a function of effect size β. The total number of phenotypes is $K = 40$, sample size is 1000, MAF is 0.3, the between-factor correlation is 0.15, and the within-factor correlation is 0.25. Significance is assessed at the 5% level.

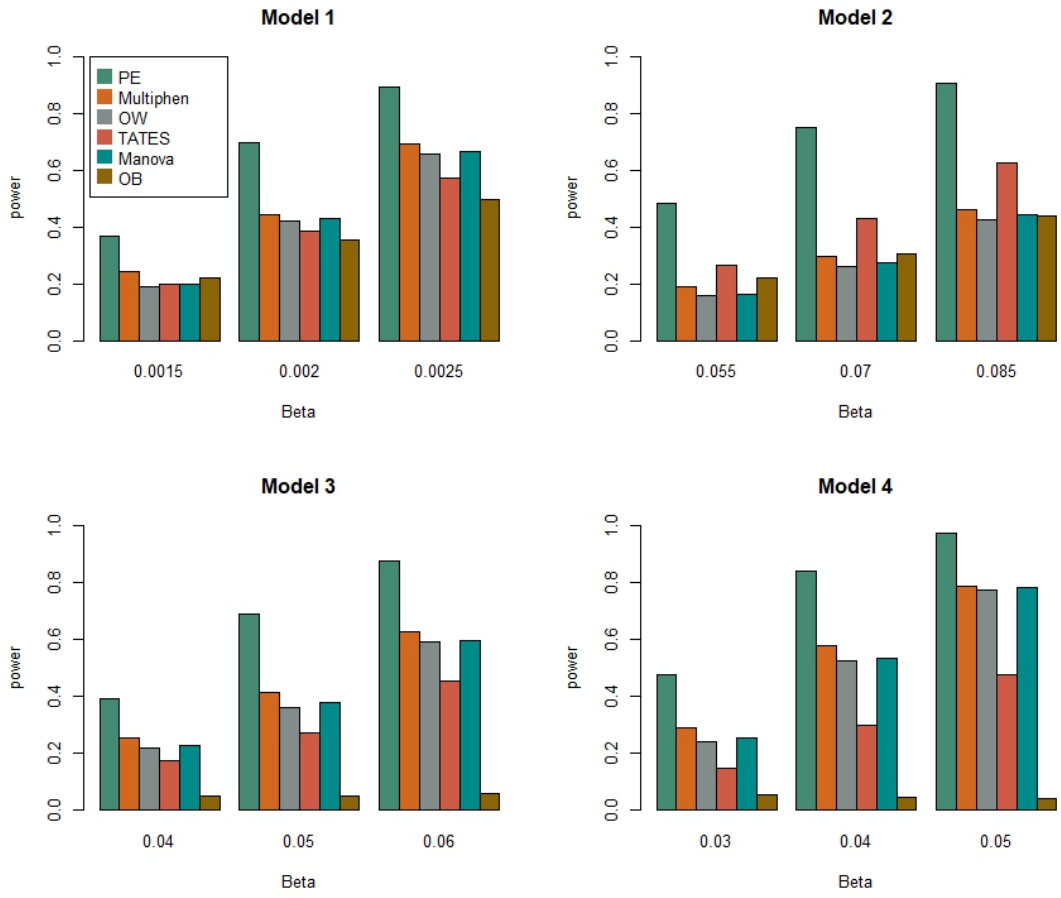


Figure 2.3 Power comparisons of the six methods as a function of within factor correlation. The total number of phenotypes is $K = 20$, sample size is 1000, MAF is 0.3, the between-factor correlation is 0.15. Significance is assessed at the 5% level.

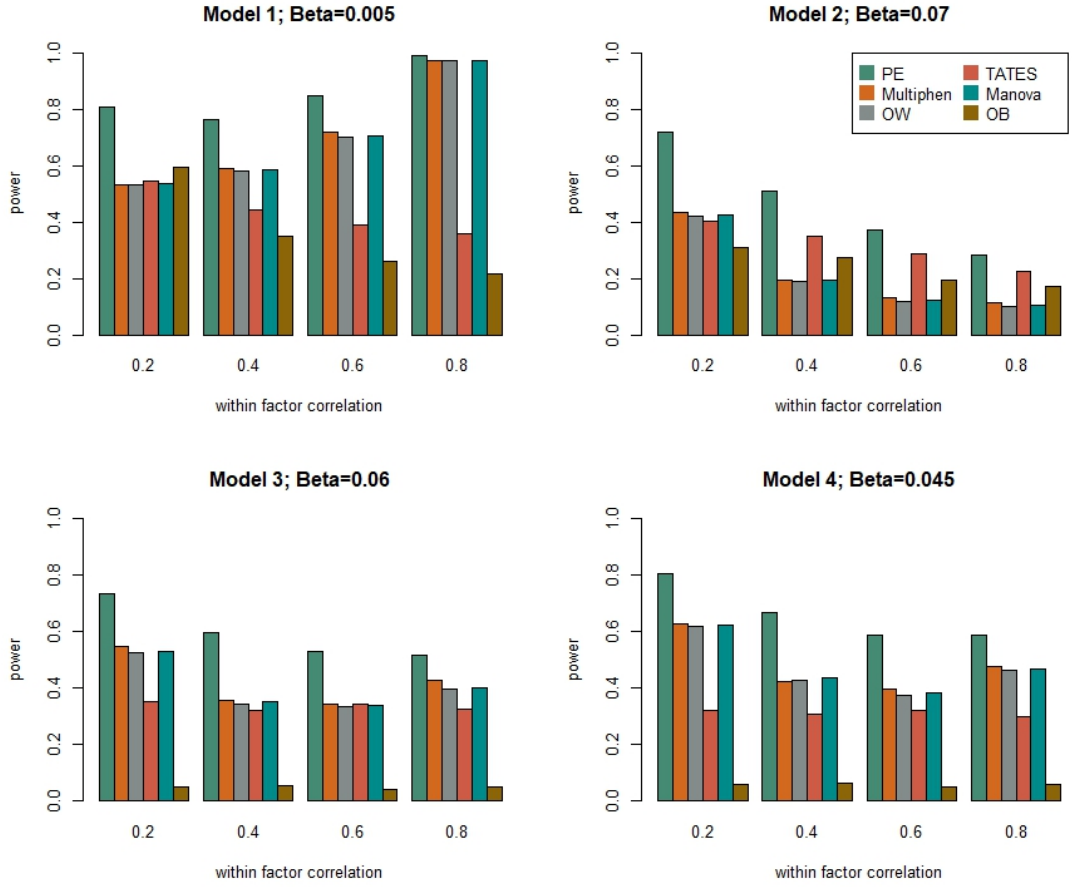
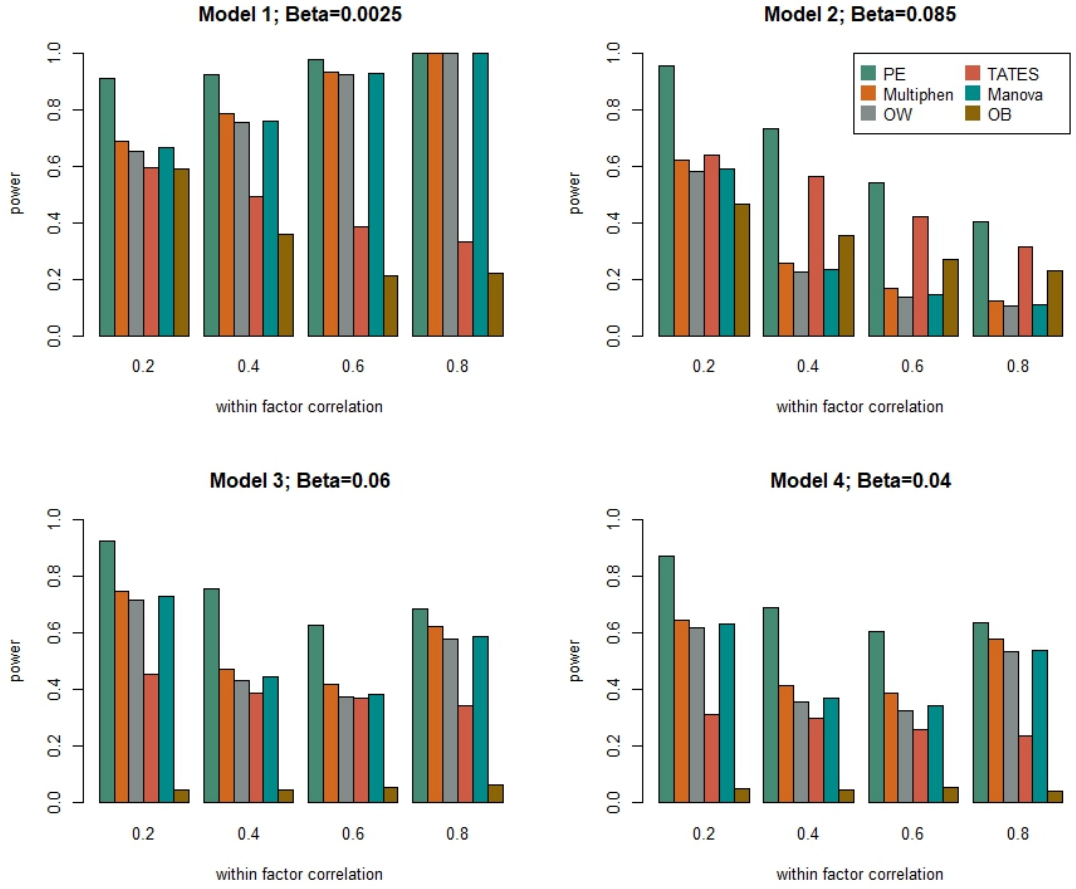


Figure 2.4 Power comparisons of the six methods as a function of within factor correlation. The total number of phenotypes is $K = 40$, sample size is 1000, MAF is 0.3, the between-factor correlation is 0.15. Significance is assessed at the 5% level.



3 Reference List

Ahituv, N., et al. (2007). "Medical Sequencing at the Extremes of Human Body Mass." The American Journal of Human Genetics **80**(4): 779-791.

Ali, A., et al. (2011). "Comparison of methods for handling missing data on immunohistochemical markers in survival analysis of breast cancer." British journal of cancer **104**(4): 693.

Allen, H. L., et al. (2010). "Hundreds of variants clustered in genomic loci and biological pathways affect human height." Nature **467**(7317): 832.

Andrés, A. M., et al. (2007). "Understanding the accuracy of statistical haplotype inference with sequence data of known phase." Genetic epidemiology **31**(7): 659-671.

Aschard, H., et al. (2014). "Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies." The American Journal of Human Genetics **94**(5): 662-676.

Ayers, K. L. and H. J. Cordell (2010). "SNP selection in genome-wide and candidate gene studies via penalized logistic regression." Genetic epidemiology **34**(8): 879-891.

Ayers, K. L. and H. J. Cordell (2013). "Identification of grouped rare and common variants via penalized logistic regression." Genetic epidemiology **37**(6): 592-602.

Bodmer, W. and C. Bonilla (2008). "Common and rare variants in multifactorial susceptibility to common diseases." Nature genetics **40**(6): 695.

Brehm, J. M., et al. (2011). "Identification of FGF7 as a novel susceptibility locus for chronic obstructive pulmonary disease." Thorax **66**(12): 1085-1090.

Cho, M. H., et al. (2010). "Variants in FAM13A are associated with chronic obstructive pulmonary disease." Nature genetics **42**(3): 200.

Cohen, J. C., et al. (2004). "Multiple rare alleles contribute to low plasma levels of HDL cholesterol." Science **305**(5685): 869-872.

Cohen, J. C., et al. (2006). "Multiple rare variants in NPC1L1 associated with reduced sterol absorption and plasma low-density lipoprotein levels." Proceedings of the National Academy of Sciences of the United States of America **103**(6): 1810-1815.

Cole, D. A., et al. (1994). "How the power of MANOVA can both increase and decrease as a function of the intercorrelations among the dependent variables." Psychological Bulletin **115**(3): 465.

Consortium, U. K. (2015). "The UK10K project identifies rare variants in health and disease." Nature **526**(7571): 82.

Chronic obstructive pulmonary disease (COPD). WHO. November 2017. Retrieved from http://www.who.int/mediacentre/factsheets/fs315/en/

Cui, K., et al. (2014). "Four SNPs in the CHRNA3/5 alpha-neuronal nicotinic acetylcholine receptor subunit locus are associated with COPD risk based on meta-analyses." PloS one **9**(7): e102324.

Cule, E. and M. De Iorio (2013). "Ridge regression in prediction problems: automatic choice of the ridge parameter." Genetic epidemiology **37**(7): 704-714.

Cule, E., et al. (2011). "Significance testing in ridge regression for genetic data." BMC bioinformatics **12**(1): 372.

De, G., et al. (2013). "Rare variant analysis for family-based design." PloS one **8**(1): e48495.

Derkach, A., et al. (2013). "Robust and powerful tests for rare variants using Fisher's method to combine evidence of association from two or more complementary tests." Genetic epidemiology **37**(1): 110-121.

Du, Y., et al. (2016). "Association of IREB2 gene rs2568494 polymorphism with risk of chronic obstructive pulmonary disease: a meta-analysis." Medical science monitor: international medical journal of experimental and clinical research **22**: 177.

Feng, S., et al. (2015). "Methods for Association Analysis and Meta-Analysis of Rare Variants in Families." Genetic epidemiology **39**(4): 227-238.

Ferreira, M. A. and S. M. Purcell (2008). "A multivariate test of association." Bioinformatics **25**(1): 132-133.

Gavish, B., et al. (2008). "Linear relationship between systolic and diastolic blood pressure monitored over 24 h: assessment and correlates." Journal of hypertension **26**(2): 199-209.

Ge, Y., et al. (2003). "Resampling-based multiple testing for microarray data analysis." Test **12**(1): 1-77.

Greco, B., et al. (2016). "A general approach for combining diverse rare variant association tests provides improved robustness across a wider range of genetic architectures." European Journal of Human Genetics **24**(5): 767.

Halawa, A. and M. El Bassiouni (2000). "Tests of regression coefficients under ridge regression models." Journal of Statistical Computation and Simulation **65**(1-4): 341-356.

Han, F. and W. Pan (2010). "A data-adaptive sum test for disease association with multiple common or rare variants." Human heredity **70**(1): 42-54.

Hancock, D. B., et al. (2010). "Meta-analyses of genome-wide association studies identify multiple loci associated with pulmonary function." Nature genetics **42**(1): 45.

Hodges, E., et al. (2007). "Genome-wide in situ exon capture for selective resequencing." Nature genetics **39**(12): 1522.

Hoerl, A. E., et al. (1975). "Ridge regression: some simulations." Communications in Statistics-Theory and Methods **4**(2): 105-123.

Hoffmann, T. J., et al. (2010). "Comprehensive approach to analyzing rare genetic variants." PloS one **5**(11): e13584.

Huang, J., et al. (2015). "Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel." Nature communications **6**: 8111.

Ionita-Laza, I., et al. (2013). "Sequence kernel association tests for the combined effect of rare and common variants." The American Journal of Human Genetics **92**(6): 841-853.

James, G., et al. (2013). An introduction to statistical learning, Springer.

Ji, W., et al. (2008). "Rare independent mutations in renal salt handling genes contribute to blood pressure variation." Nature genetics **40**(5): 592.

Kim, J., et al. (2015). "An adaptive association test for multiple phenotypes with GWAS summary statistics." Genetic epidemiology **39**(8): 651-663.

Klei, L., et al. (2008). "Pleiotropy and principal components of heritability combine to increase power for association analysis." Genetic epidemiology **32**(1): 9-19.

Lange, L. A., et al. (2014). "Whole-exome sequencing identifies rare and low-frequency coding variants associated with LDL cholesterol." The American Journal of Human Genetics **94**(2): 233-245.

Lee, S., et al. (2012). "Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies." The American Journal of Human Genetics **91**(2): 224-237.

Li, B. and S. M. Leal (2008). "Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data." The American Journal of Human Genetics **83**(3): 311-321.

Li, X., et al. (2011). "Importance of hedgehog interacting protein and other lung function genes in asthma." Journal of Allergy and Clinical Immunology **127**(6): 1457-1465.

Liang, X., et al. (2016). "An Adaptive Fisher's Combination Method for Joint Analysis of Multiple Phenotypes in Association Studies." Scientific Reports **6**: 34323.

Lin, D.-Y. and Z.-Z. Tang (2011). "A general framework for detecting disease associations with rare variants in sequencing studies." The American Journal of Human Genetics **89**(3): 354-367.

Locke, D. E., et al. (2006). "Relationship of indicators of neuropathology, psychopathology, and effort to neuropsychological results in patients with

epilepsy or psychogenic non-epileptic seizures." Journal of clinical and experimental neuropsychology **28**(3): 325-340.

Luntz, A. (1969). "On estimation of characters obtained in statistical procedure of recognition." Technicheskaya Kibernetica **3**.

Lutz, S. M., et al. (2015). "A genome-wide association study identifies risk loci for spirometric measures among smokers of European and African ancestry." BMC genetics **16**(1): 138.

Madsen, B. E. and S. R. Browning (2009). "A groupwise association test for rare mutations using a weighted sum statistic." PLoS genetics **5**(2): e1000384.

Malo, N., et al. (2008). "Accommodating linkage disequilibrium in genetic-association analyses via ridge regression." The American Journal of Human Genetics **82**(2): 375-385.

Manolio, T. A., et al. (2009). "Finding the missing heritability of complex diseases." Nature **461**(7265): 747.

Marini, N. J., et al. (2008). "The prevalence of folate-remedial MTHFR enzyme variants in humans." Proceedings of the National Academy of Sciences **105**(23): 8055-8060.

McCarthy, M. I., et al. (2008). "Genome-wide association studies for complex traits: consensus, uncertainty and challenges." Nature reviews genetics **9**(5): 356.

Meier, L., et al. (2008). "The group lasso for logistic regression." Journal of the Royal Statistical Society: Series B (Statistical Methodology) **70**(1): 53-71.

Morgenthaler, S. and W. G. Thilly (2007). "A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST)." Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis **615**(1): 28-56.

Ng, S. B., et al. (2009). "Targeted capture and massively parallel sequencing of 12 human exomes." Nature **461**(7261): 272.

O'Brien, P. C. (1984). "Procedures for comparing samples with multiple endpoints." Biometrics: 1079-1087.

O'Reilly, P. F., et al. (2012). "MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS." PloS one **7**(5): e34861.

Pillai, S. G., et al. (2009). "A genome-wide association study in chronic obstructive pulmonary disease (COPD): identification of two major susceptibility loci." PLoS genetics **5**(3): e1000421.

Price, A. L., et al. (2010). "Pooled association tests for rare variants in exon-resequencing studies." The American Journal of Human Genetics **86**(6): 832-838.

Price, A. L., et al. (2006). "Principal components analysis corrects for stratification in genome-wide association studies." Nature genetics **38**(8): 904.

Pritchard, J. K. (2001). "Are rare variants responsible for susceptibility to complex diseases?" The American Journal of Human Genetics **69**(1): 124-137.

Pritchard, J. K. and N. J. Cox (2002). "The allelic architecture of human disease genes: common disease–common variant… or not?" Human molecular genetics **11**(20): 2417-2423.

Ray, D., et al. (2016). "USAT: A Unified Score-Based Association Test for Multiple Phenotype-Genotype Analysis." Genetic epidemiology **40**(1): 20-34.

Romeo, S., et al. (2007). "Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL." Nature genetics **39**(4): 513.

Romeo, S., et al. (2009). "Rare loss-of-function mutations in ANGPTL family members contribute to plasma triglyceride levels in humans." The Journal of clinical investigation **119**(1): 70-79.

Sha, Q., et al. (2013). "Adaptive clustering and adaptive weighting methods to detect disease associated rare variants." European Journal of Human Genetics **21**(3): 332.

Sha, Q., et al. (2012). "Detecting association of rare and common variants by testing an optimally weighted combination of variants." Genetic epidemiology **36**(6): 561-571.

Sha, Q. and S. Zhang (2014). "A Rare Variant Association Test Based on Combinations of Single-Variant Tests." Genetic epidemiology **38**(6): 494-501.

Shen, L., et al. (2010). "Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in MCI and AD: A study of the ADNI cohort." Neuroimage **53**(3): 1051-1063.

Stratton, M. R. and N. Rahman (2008). "The emerging landscape of breast cancer susceptibility." Nature genetics **40**(1): 17.

Sun, J., et al. (2016). "A method for analyzing multiple continuous phenotypes in rare variant association studies allowing for flexible correlations in variant effects." European Journal of Human Genetics **24**(9): 1344.

Taylor, P. N., et al. (2015). "Whole-genome sequence-based analysis of thyroid function." Nature communications **6**: 5681.

Teer, J. K. and J. C. Mullikin (2010). "Exome sequencing: the sweet spot before whole genomes." Human molecular genetics **19**(R2): R145-R151.

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso." Journal of the Royal Statistical Society. Series B (Methodological): 267-288.

Van der Sluis, S., et al. (2013). "TATES: efficient multivariate genotype-phenotype analysis for genome-wide association studies." PLoS genetics **9**(1): e1003235.

Walsh, T. and M.-C. King (2007). "Ten genes for inherited breast cancer." Cancer cell **11**(2): 103-105.

Wang, Y., et al. (2015). "Pleiotropy analysis of quantitative traits at gene level by multivariate functional linear models." Genetic epidemiology **39**(4): 259-275.

Wang, Z., et al. (2016). "Joint analysis of multiple traits using" optimal" maximum heritability test." PloS one **11**(3): e0150975.

Warren, H., et al. (2014). "Genetic prediction of quantitative lipid traits: comparing shrinkage models to gene scores." Genetic epidemiology **38**(1): 72-83.

Wilk, J. B., et al. (2009). "A genome-wide association study of pulmonary function measures in the Framingham Heart Study." PLoS genetics **5**(3): e1000429.

Wilk, J. B., et al. (2012). "Genome-wide association studies identify CHRNA5/3 and HTR4 in the development of airflow obstruction." American journal of respiratory and critical care medicine **186**(7): 622-632.

Wu, B. and J. S. Pankow (2016). "Sequence kernel association test of multiple continuous phenotypes." Genetic epidemiology **40**(2): 91-100.

Wu, M. C., et al. (2011). "Rare-variant association testing for sequencing data with the sequence kernel association test." The American Journal of Human Genetics **89**(1): 82-93.

Yang, J. J., et al. (2016). "An efficient genome-wide association test for multivariate phenotypes based on the Fisher combination function." BMC bioinformatics **17**(1): 19.

Yang, Q., et al. (2010). "Analyze multivariate phenotypes in genetic association studies by combining univariate association tests." Genetic epidemiology **34**(5): 444-454.

Yang, X., et al. (2017). "Detecting association of rare and common variants based on cross-validation prediction error." Genetic epidemiology **41**(3): 233-243.

Yi, N. and D. Zhi (2011). "Bayesian analysis of rare variants in genetic association studies." Genetic epidemiology **35**(1): 57-69.

Young, R., et al. (2010). "Chromosome 4q31 locus in COPD is also associated with lung cancer." European Respiratory Journal **36**(6): 1375-1382.

Yuan, M. and Y. Lin (2006). "Model selection and estimation in regression with grouped variables." Journal of the Royal Statistical Society: Series B (Statistical Methodology) **68**(1): 49-67.

Zhang, J., et al. (2011). "Nicotinic acetylcholine receptor variants associated with susceptibility to chronic obstructive pulmonary disease: a meta-analysis." Respiratory research **12**(1): 158.

Zheng, H. F., et al. (2015). "Whole-genome sequencing identifies EN1 as a determinant of bone density and fracture." Nature **526**(7571): 112.

Zhu, A. Z., et al. (2014). "Association of CHRNA5-A3-B4 SNP rs2036527 With Smoking Cessation Therapy Response in African-American Smokers." Clinical Pharmacology & Therapeutics **96**(2): 256-265.

Zhu, H., et al. (2015). "Power comparisons of methods for joint association analysis of multiple phenotypes." Human heredity **80**(3): 144-152.

Zhu, H., et al. (2018). "A novel method to test associations between a weighted combination of phenotypes and genetic variants." PloS one **13**(1): e0190788.

Zhu, X., et al. (2010). "Detecting rare variants for complex traits using family and unrelated data." Genetic epidemiology **34**(2): 171-187.

Appendix A: The closed-form formula of cross-validation prediction error of LOOCV for Ridge regression

Let $x_i = \left(1, g_{i1}, \ldots g_{iM}\right)^T$ and $X = \left(x_1, \ldots, x_n\right)^T$. Let $A_\lambda = \left(X^T X + \lambda I\right)^{-1}$, $h_i^\lambda = x_i^T A_\lambda x_i$, and $\hat{\beta}_\lambda = A_\lambda X^T y$. Let $\hat{y}_i^\lambda = x_i^T \hat{\beta}_\lambda$ and $B_\lambda = X A_\lambda X^T$, then $h_\lambda = \left(h_1^\lambda, ..., h_n^\lambda\right) = diag\left(B_\lambda\right)$. Let X_{-i}, $\hat{\beta}_{\lambda,-i}$, and $\hat{y}_{ci}^\lambda$ denote X, $\hat{\beta}_\lambda$, and $\hat{y}_i^\lambda$ when the i^{th} individual leaves out. Noting that $X_{-i}^T X_{-i} = X^T X - x_i x_i^T$, then we have

$$A_{\lambda,-i} = \left(X^T X + \lambda I - x_i x_i^T\right)^{-1} = A_\lambda + \frac{1}{1 - x_i^T A_\lambda x_i} A_\lambda x_i x_i^T A_\lambda,$$

$$\hat{\beta}_{\lambda,-i} = A_{\lambda,-i} X_{-i}^T y_{-i} = \left(A_\lambda + \frac{1}{1 - a_i^\lambda} A_\lambda x_i x_i^T A_\lambda\right)\left(X^T y - x_i y_i\right)$$

$$= A_\lambda X^T y - A_\lambda x_i y_i + \frac{1}{1 - h_i^\lambda} A_\lambda x_i x_i^T A_\lambda X^T y - \frac{h_i^\lambda}{1 - h_i^\lambda} A_\lambda x_i y_i,$$

$$\hat{y}_{ci}^\lambda = x_i^T \hat{\beta}_{\lambda,-i} = x_i^T \left(A_\lambda X^T y - A_\lambda x_i y_i + \frac{1}{1 - h_i^\lambda} A_\lambda x_i x_i^T A_\lambda X^T y - \frac{h_i^\lambda}{1 - h_i^\lambda} A_\lambda x_i y_i\right)$$

$$= \hat{y}_i^\lambda - h_i^\lambda y_i + \frac{h_i^\lambda}{1 - h_i^\lambda} \hat{y}_i^\lambda - \frac{\left(h_i^\lambda\right)^2}{1 - h_i^\lambda} y_i = \frac{1}{1 - h_i^\lambda} \hat{y}_i^\lambda - \frac{h_i^\lambda}{1 - h_i^\lambda} y_i.$$

Therefore, $y_i - \hat{y}_{ci}^\lambda = \frac{1}{1 - h_i^\lambda}\left(y_i - \hat{y}_i^\lambda\right)$.

Appendix B: The fast algorithms for permutation procedures

1) PE method

We use the same notations as in Appendix A. Let $\hat{y}_\lambda = \left(\hat{y}_1^\lambda, \ldots, \hat{y}_n^\lambda\right)^T$ and then $\hat{y}_\lambda = X\left(X^T X + \lambda I\right)^{-1} X^T y$. For two matrices or vectors A and B, we use $A * B$ and $\frac{A}{B}$ to denote the element-wise operations. Let m denote the number of columns of matrix X. We assume $n \geq m$. We perform singular value decomposition of X, that is, $X = UDV$, where U is an $n \times m$ matrix with orthonormal columns, D is $m \times m$ diagonal matrix with non-negative real numbers on the diagonal, and V is an $m \times m$ orthogonal matrix. Let $D = diag\left(d_1, \ldots, d_m\right)$. Then, $\hat{y}_\lambda = UC_\lambda U^T y$, where $C_\lambda = diag\left(c_{\lambda,1}, \ldots, c_{\lambda,m}\right)$ and $c_{\lambda,j} = \frac{d_j^2}{d_j^2 + \lambda}$ for $j = 1, \ldots, m$. Let $c_\lambda = \left(c_{\lambda,1}, \ldots, c_{\lambda,m}\right)^T$ and $y^{(m)} = U^T y$ be a m dimensional vector. Then, $\hat{y}_\lambda = UC_\lambda y^{(m)} = U\left(c_\lambda * y^{(m)}\right)$ and $h_\lambda = diag\left(UC_\lambda U^T\right)$ (in R code, $h_\lambda = rowSums\left(U * t\left(t\left(U\right) * c_\lambda\right)\right)$). For $0 \leq \lambda_1 < \ldots < \lambda_{K-1} < \lambda_K < \infty$, let $C = \left(c_{\lambda_1}, \ldots, c_{\lambda_K}\right)$ and $H = \left(h_{\lambda_1}, \ldots, h_{\lambda_K}\right)$. Then, $\left(\hat{y}_{\lambda_1}, \ldots, \hat{y}_{\lambda_K}\right) = U\left(C * y^{(m)}\right) = U\left(c_{\lambda_1} * y^{(m)}, \ldots, c_{\lambda_K} * y^{(m)}\right)$. If we denote $B = \frac{\left(y - \hat{y}_{\lambda_1}, \ldots, y - \hat{y}_{\lambda_K}\right)}{1 - H}$, then $\left(T_{\lambda_1}, \ldots, T_{\lambda_K}\right) = colSums\left(B * B\right)$. Note that C, U, and H do not change in each permutation.

2) PE-TOW and PE-WS methods

For PE-TOW, let $X = (x_1, \ldots, x_n)^T$, where $x_i = G_i - \bar{G}$. We first centralize the trait values y. For simplicity, we still use $y = (y_1, \ldots, y_n)^T$ to denote the trait values after centralization. Let $d = X^T X = \sum_{i=1}^{n} x_i^2$, $Y = X^T y = \sum_{i=1}^{n} x_i y_i$, and $c_\lambda = \frac{1}{d + \lambda}$. Then, $\hat{y}_\lambda = Y c_\lambda X$ and $h_i^\lambda = c_\lambda x_i^2$. Note that we need to recalculate X in each permutation.

For PE-WS, the same formulas for PE-TOW are applied. However, X does not change in each permutation.